www.sciencemag.org/cgi/content/full/science.1196914/DC1

Supporting Online Material for

Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project

Mark B. Gerstein,* Zhi John Lu, Eric L. Van Nostrand, Chao Cheng, Bradley I. Arshinoff, Tao Liu, Kevin Y. Yip, Rebecca Robilotto, Andreas Rechtsteiner, Kohta Ikegami, Pedro Alves, Aurelien Chateigner, Marc Perry, Mitzi Morris, Raymond K. Auerbach, Xin Feng, Jing Leng, Anne Vielle, Wei Niu, Kahn Rhrissorrakrai, Ashish Agarwal, Roger P. Alexander, Galt Barber, Cathleen M. Brdlik, Jennifer Brennan, Jeremy Jean Brouillet, Adrian Carr, Ming-Sin Cheung, Hiram Clawson, Sergio Contrino, Luke O. Dannenberg, Abby F. Dernburg, Arshad Desai, Lindsay Dick, Andréa C. Dosé, Jiang Du, Thea Egelhofer, Sevinc Ercan, Ghia Euskirchen, Brent Ewing, Elise A. Feingold, Reto Gassman, Peter J. Good, Phil Green, Francois Gullier, Michelle Gutwein, Mark S. Guyer, Lukas Habegger, Ting Han, Jorja G. Henikoff, Stefan R. Henz, Angie Hinrichs, Heather Holster, Tony Hyman, A. Leo Iniguez, Judith Janette, Morten Jensen, Masaomi Kato, W. James Kent, Ellen Kephart, Vishal Khivansara, Ekta Khurana, John K. Kim, Paulina Kolasinska-Zwierz, Eric C. Lai, Isabel Latorre, Amber Leahey, Suzanna Lewis, Paul Lloyd, Lucas Lochovsky, Rebecca F. Lowdon, Yaniv Lubling, Rachel Lyne, Michael MacCoss, Sebastian D. Mackowiak, Marco Mangone, Sheldon McKay, Desirea Mecenas, Gennifer Merrihew, David M. Miller III, Andrew Muroyama, John I. Murray, Siew-Loon Ooi, Hoang Pham, Taryn Phippen, Elicia A. Preston, Nikolaus Rajewsky, Gunnar Rätsch, Heidi Rosenbaum, Joel Rozowsky, Kim Rutherford, Peter Ruzanov, Mihail Sarov, Rajkumar Sasidharan, Andrea Sboner, Paul Scheid, Eran Segal, Hyunjin Shin, Chong Shou, Frank J. Slack, Cindie Slightam, Richard Smith, William C. Spencer, E.O. Stinson, Scott Taing, Teruaki Takasaki, Dionne Vafeados, Ksenia Voronina, Guilin Wang, Nicole L. Washington, Christina Whittle, Beijing Wu, Koon-Kiu Yan, Georg Zeller, Zheng Zha, Mei Zhong, Xingliang Zhou, modENCODE Consortium, Julie Ahringer,* Susan Strome,* Kristin C. Gunsalus,* Gos Micklem,* X. Shirley Liu,* Valerie Reinke,* Stuart K. Kim,* LaDeana W. Hillier,* Steven Henikoff,* Fabio Piano,* Michael Snyder,* Lincoln Stein,* Jason D. Lieb,* Robert H. Waterston*

*To whom correspondence should be addressed. E-mail: modencode.worm.pi@gersteinlab.org

Published 22 December 2010 on *Science* Express
DOI: 10.1126/science.1196914

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1196914/DC1)

Table S18 as a zipped Excel file

Supplementary Text for “Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project”

A. Overview of the Supplement and Online Resources

A.1. The Supplement

In the supplement we provide more details on the data and analysis described in the main text. Below is an outline of the major sections of the document to give an overview of key components. Note that the supplement is laid out in a parallel fashion to the main text, as much as possible sharing common headings. Where an outline heading is exactly parallel, it is prefixed by "More Detail on" and then has its section name from the main text “quoted and underlined”.

A.2. The Paper Site: modencode.org/publications/integrative_worm_2010

All modENCODE data and analyses are available online. To explore the underlying details of the datasets specifically presented in this paper further, we recommend as a first point of reference utilizing the paper web site at http://www.modencode.org/publications/integrative_worm_2010. The site serves as a central resource point for accessing all data associated with this paper. On this page, for all figures and tables presented (including those in the supplement), we have listed links to the underlying source data (for each of the experiments analyzed) and any intermediate analysis files that aggregate the source data in various ways or reference external data sources. We have also included links to many of the tools used in the analysis. Finally, we have stated the WormBase version(s) to which data has been mapped or compared for each section in Table S17. The list of experiment links has descriptive titles, making it easy to identify and access individual experiments analyzed using the modMine interface (*1*).

A.3. modENCODE.org

The modENCODE project website, www.modENCODE.org, is the primary entry point for accessing and downloading the entire modENCODE data corpus.

Following the modMine link from the modencode.org provides a searchable interface and easy to explore organization of the datasets. For access to a graphical depiction of the datasets across the chromosomes, follow the "Browse worm Genomes" link to open a GBrowser window The GBrowser enables side by side visual comparison of datasets and provides options to customize, share and export regions of interest.

A.4. WormBase, SRA and Beyond

Finally, ModENCODE data and analyses are available through many international repositories in various forms. The primary site to access and download the six-way nematode alignment is the UCSC Genome Browser (*2*), and raw microarray and sequencing data are available from the GEO (*3*) and SRA (*4*) resources respectively. The accession numbers for GEO and SRA data sets can be found linked from the modMine dataset summaries, or the resources can be searched directly for the "modENCODE" project. Interpreted data, including corrected gene models, alternative transcripts, and ChIP peaks, are being incorporated into WormBase (*5*). Interested users can also apply for access to the Bionimbus private compute cloud, an experimental resource that holds a complete mirror of the modENCODE corpus and virtual machines that are pre-populated with a variety of tools for accessing and manipulating the data.

B. Information on the Initial Data Preparation

B.1. Information on Embryo Staging

Here we discuss how embryos were staged for these studies. A previous study hand-selected embryos and then performed two rounds of amplification in order to analyze gene expression profiles (*6*). They were therefore able to precisely stage the embryos to a specific number of cells, and perform a high-resolution timecourse as embryos progressed from 1- to 2- to 4- to 8-cell embryos. We required a much greater amount of starting material and chose not to use amplification, so we did not handpick our embryos but instead collected embryos directly from young adults by bleaching. Embryos collected immediately for analysis (early stage, or EE embryos) consisted on average of 37% <28-cell embryos, 30% 28–100-cell embryos, and 33% ~100–300-cell embryos. Embryos were also allowed to progress through development for ~6 hours (late stage, or LE embryos), and then were collected. These late stage embryos consisted of a diverse mixture of embryos in the comma, one-fold, and two-fold stages.

B.2. Minimizing Batch Effects in Sample Preparation

Batch effects are an important source of error that can confound analysis of high-throughput functional genomics data (*7*). We have taken a number of steps to measure them and ameliorate their effect.

We have tried where possible to centralize the sample preparation. All the samples for the tiling arrays were done in the same lab and these were the same as for the RNA-seq. In particular, the samples for the transcriptome analysis were generated in the labs of Reinke (whole animal) and Miller (tissue-specific). These two labs coordinated sample preparation as much as possible, given the different experimental constraints of their approaches. Animals were synchronized to a two-hour window in the early larval stage, and then growth was timed to each of the subsequent developmental stages.

For the whole animal samples, RNA was collected using the Trizol method (*8*) and directly analyzed by tiling array or RNA-seq. All tiling array hybridizations were performed by the same person at a core facility at Yale. The RNA-seq samples for each stage were performed on the exact same RNA population used for the tiling array in most cases, with a few instances of having to re-isolate the RNA from a independently grown prep, which underwent exactly the same synchronization procedure and growth conditions. Correlation analysis between tiling array and RNA-seq for these independent samples indicates that they are nearly as closely related as when the same RNA sample is analyzed by both tiling array and RNA-seq. The Spearman correlation coefficients relating RNA-seq vs. tiling array data for late embryo (independent sample preparations) and L2 (same sample preparation) are 0.85 and 0.82, respectively.

For the tissue-specific samples (on which only tiling array analysis was performed), all samples were compared to an internally created control sample, which represents the whole animal at that particular stage. Given the specific experimental manipulation (FACS sorting or IP) required to generate the tissue-specific samples, it was more appropriate to use this internal control, which had undergone the same manipulations, rather than those generated in another lab. All tiling array hybridizations for tissue-specific samples and controls were performed in the Vanderbilt microarray core facility. Moreover, all of these samples were hybridized to arrays in triplicate, from three independently grown and isolated preps, demonstrating reproducibility.

B.3. Comparing and Scaling Array and Sequencing Data

B.3.a. ChIP-chip vs. ChIP-seq

The modENCODE project began when tiling arrays (*9*) were still the platform of choice for genome-wide location analysis. Many genome-wide location data sets, especially on histone marks and chromatin factors, were obtained using ChIP-chip (*10*) on tiling arrays. To ensure the compatibility between ChIP-chip and ChIP-seq data generated by different modENCODE groups, we examined RNA Pol II ChIP data detected by both ChIP-chip (from the Lieb project) and ChIP-seq (from the Snyder project) (Fig. S1).

At 1 kb resolution, the correlation between individual RNA Pol II profiles at a given stage is 0.75-0.88 within ChIP-seq replicates and 0.77-0.91 within ChIP-chip replicates. The correlation scores between ChIP-seq and ChIP-chip replicates are 0.56-0.78. Although variations across platform/group are slightly higher than those within platform/group, data across different labs at the same stage are still more correlated than those across different stages by the same lab.

Finally, while ChIP-seq yielded more peaks than did ChIP-chip, the top 3,000 peaks identified by ChIP-chip and ChIP-seq overlap by approximately 2/3, a level of agreement normally observed for ChIP data from different labs on the same platform. These observations not only indicate that the two platforms are comparable, but also attest to the high quality of the respective data sets.

B.3.b. Expression Tiling Arrays vs. RNA-seq

We also had an opportunity to compare tiling array and RNA-seq technologies for measurement of gene expression, as data sets were generated using both techniques on matched samples. As there were no biological replicates in the RNA-seq time course, we made use of the fact that for many RNA-seq experiments, the identical RNA source was profiled by tiling array. This allowed us to perform comparisons of the data generated by the two methods. A detailed comparison of these methods was described in (*11*); in addition to presenting some main points from this analysis here, we also repeat this analysis on data sets associated with this manuscript. From this comparison, we were able to develop methods of optimally scaling the tiling array measurements to make them best correspond to those from RNA-seq.

Signals from the two platforms agree well (Fig. S2). For a young adult sample, the Pearson correlation is 0.83 between RNA-seq measurements using polyA-selected RNA and tiling array measurements using total RNA. A higher correlation of 0.90 was found when polyA-enrichment was also used for the sample that had been hybridized on tiling arrays. Using the maxgap-minrun algorithm with optimized parameters, we then segmented the signals into transcriptionally active regions (*12, 13*). A ROC curve, parameterized by signal threshold, indicates that RNA-seq consistently outperforms tiling array in its ability to predict known transcribed regions. For instance, at a false positive rate (FPR) of 0.05, the tiling array yields a sensitivity of 0.68, while RNA-seq attains a sensitivity of 0.85. Correspondingly, we also found that the RNA-seq data predicted exon boundaries with greater accuracy, with a median offset of 0 bp (in comparison to 7 bp for the tiling array data). This is to be expected, as the resolution of an array is limited by its probe size, which was 25 bp in this experiment.

Fig. S2 shows several genes in the upper left, indicating they are measured as highly expressed by tiling array but not RNA-seq. We conducted a "nearest neighbor" analysis to investigate the hypothesis that this is due to cross-hybridization effects on the array. For each gene, we computed the expression level from probes lying within that gene, as well as probes similar in sequence, but elsewhere in the genome. For tiling arrays, we found these two values to be similar for many genes, indicating that the suggested expression could arise equally well from true expression or cross-hybridization. These values are similar for fewer genes when using RNA-seq data. Another analysis, using pseudogenes, also confirms cross-hybridization in arrays (*11*). We have used these analyses in formulating our fairly conservative criteria for transcribed pseudogenes (see main text and Fig. 1D).

For determining gene expression values maximally compatible with RNA-seq, we used the following procedure: for 42 of the 46 experiments listed in Table S3 (without some of the infection samples), we obtained a signal track by applying pseudomedian smoothing over the three replicates, which provides an expression level for each probe. We then consider all probes overlapping the exonic regions of each transcript by at least 50%. We defined the expression level of this transcript as the median of the signal values for all such probes. Gene expression levels were then defined simply as equal to those of the longest isoform. For the inter-sample comparison, we normalized these expression levels by dividing the values by the slide median, i.e. the median of all probes on the array and obtained a large data matrix (42 samples x 20,085 genes). Expression levels for each slide were next centered by subtracting the mean expression value for each slide from all expression values within the slide.

C. More Detail on "the Transcriptome"

C.1. More Detail on "Protein-coding Genes" and "Gene Models"

C.1.a. RNA-seq Read Mapping and Methods for the Creation of Stage-Specific RNA-seq-only Genelets

Stage-specific genelets, based solely on stage-specific RNA-seq data were created using methods similar to those described (*14*), but with several additional refinements. Briefly, the Illumina reads were uniquely aligned against the genome, and an exhaustive coverage-based spliced leader and splice junction database were created for each stage (*14*). Thresholds for read coverage were set for a 0.05 false positive rate, based on a ROC analysis. Transcripts were created by seeding with the highest confidence splice sites and spliced leaders in a region, and then extending from those sites and leaders, incorporating coverage and junctions into the model (Fig. S5). The procedure was iterated until all confirmed splice junctions and leaders were incorporated into models. Instead of producing transcripts containing every possible combination of every splice junction/leader, each splice junction/leader was used in at least one model. We created alternative models, with merged neighboring exons, when above-threshold read coverage suggested the intron had been retained, and when frame was maintained across the merged region. We also generated genelets with alternative start/stop sites within introns when the entire intron was not retained, but when there were at least 50 bases of above-threshold coverage that extended into the intron initiated by a TSS or terminated by a polyA site.

To generate the list of polyA addition sites for possible inclusion in our transcript sets, we first created a list of all possible blocks that could contain a polyA site by using a non-redundant list of exons from our integrated transcript set and identifying all blocks between the start of one exon and the start of the next upstream exon. For each of those blocks, we took 3P-Seq tags from (*15*) (defined in our analysis as the tags with at least one 3´-terminal A, at least one of which was untemplated) and found the site with the highest number of tags in that interval and clustered the tags +/-10 bases from that site. We then looked for the site with the next highest number of tags and clustered around that site, demanding that the cluster have at least five tags and have at least 1% of the tags in the first cluster. We continued in that manner to identify all candidate sites. For all sites in that block that had less than five tags or had <1% as many reads as the most supported candidate site, we labeled them as "secondary". For blocks where there were no sites with >=5 tags, all sites within those blocks were labeled as "orphans". For secondary and orphan sites, we retained the site if a polyA site defined by another method (Mangone et al., 2010, RNA-seq or WormBase) was present within a distance 10 nucleotides of the site. Some sites defined by our RNA-seq data, Mangone et al., or WormBase fell >10 nucleotides away from any 3P-Seq tag; these also were retained as candidate polyA sites. The final set of polyA sites (30,737) was then the subset of the candidate sites where, during transcript creation, above-threshold coverage extended to reach the site. For these 30,737 sites, we tallied those that had support from each data source, considering a query site supported if the source indicated that a site was present within 20 nucleotides of the query site (Table S2a).

The stage-specific polyA addition sites (including those generated by this project, as well as those from (*15*)) were clustered (keeping only a single polyA addition site when there are multiple polyA sites within a 10 bp window). While all spliced leaders were incorporated into at least one prediction, polyA sites were only incorporated when a genelet model extended to the polyA site. Because overlapping UTRs can cause neighboring same strand predictions to merge if there is no spliced leader or no polyA site, whenever a single exon overlapped two separate neighboring WormBase gene predictions, we broke the corresponding transcript into two separate transcripts. We also broke transcripts whenever they overlapped more than one WormBase gene prediction, and three or more neighboring exons were not included in the CDS portion of the transcript. The CDS region was defined by identifying the longest open reading frame. Single exon transcripts from WormBase were incorporated if at least 75 bases had above-threshold coverage. Additionally, single exon transcripts were created when a single block of coverage was at least 75 bases long and extended from an SL to a polyA site, or if it began with an SL and extended at least 250 bases (even if without a polyA site).

C.1.b. Methods for the Aggregate Integrated Transcript Set

To create the aggregate integrated transcript set, all of the reads (from all stages) were combined as if they were from a "single project". Splice junctions, spliced leaders, and polyA addition sites were identified as they would be in the stage-specific methods. Transcripts were then built in the way described above, seeding with splice junctions and extending using "experimentally confirmed" bases (see below). However, additional evidence from mRNAs/ESTs, WormBase, and modENCODE data were incorporated as described here.

The following splice junctions were included in the aggregate integrated set: (1) splice junctions confirmed in the individual RNA-seq stages or by aggregate read coverage, (2) splice junctions confirmed by mRNA/EST in WormBase (WS209), RT-PCR/RACE, and mass spectrometry (*16*), and (3) WormBase-predicted splice junctions which were supported by RNA-seq data (including those after allowing an RNA-seq read to be placed in all positions at which it had an identical match). Note that, for the splice junction counts in Fig. 1A, we counted any splice junction beginning “before” the 5’ end of an existing WormBase (WS170) transcript prediction as 5’. Similarly, any splice junction extending “past” the 3’ end of an existing WormBase transcript prediction was annotated as 3’. Any splice junction internal to a WormBase transcript prediction was labeled as internal.

In the aggregate transcript set, a base was considered experimentally confirmed when any one of the following criteria were met: (a) above-threshold coverage in the individual stage or aggregate RNA-seq data set, (b) coverage by an mRNA/EST, RT-PCR/RACE, or alignment by mass spectrometry, (c) coverage by WormBase predictions, as long as the bounding splice junctions are confirmed splice junctions (i.e. holes in coverage within exons which already have evidence based on RT-PCR, RNA-seq, EST/mRNA, etc. can be "filled in" using WormBase coverage), or (d) coverage by a genelet created in the individual stage-specific sets. In addition to the "integrated transcript set," we also created an "integrated genelet set" where evidence "(c)", supplemented with WormBase predictions, is not included.

For the aggregate set, spliced leader and polyA addition site data were included when (a) they were defined by coverage in individual stages (novel spliced leaders or polyA sites defined by the RNA-seq-only analyses were required to appear in more than one of the individual stages to be included) and/or in the aggregate set, (b) they were identified by WormBase (WS209) as SLs or polyAs, (c) they were identified in other studies generated from deep 3' RACE sequencing (*15, 17*), or (d) spliced leaders were identified by RT-PCR/RACE experiments.

Transcripts are named after the overlapping WormBase transcript. For instance, the alternative transcripts/isoforms associated with WormBase C10H11.1 would have names such as C10H11.1.T1, C10H11.1.T2, C10H11.1.T3, etc. Those transcripts which do not overlap a WormBase transcript have names beginning with "RIT*" (for RNA-seq Integrated Transcript). The number following "RIT" is the chromosome (1=I, 2=II, etc. 6=X). The number after the first period is a unique number assigned to that transcript. The T1, T2, etc. are used for the alternative versions of that transcript. Currently, the naming does not allow one to know which transcript versions have the same CDS.

For the aggregate transcript set described here, we included all of the 19 stages for which RNA-seq data was available (Fig. S3).

C.1.c. RNA-seq Saturation Analysis

In order to understand the relationship between the robustness of gene expression measurements and the depth of sequencing, we devised the following *in silico* experiments:

1. We considered an RNA-seq experiment with ~36M mapped reads (mid-L2 25dC 14 hours post L1 - DCCid=2351);
2. We randomly selected fractions of the mapped reads: 1%, 5%, 10%, …, 90%, such that we generated subsets of ~300K, 1.6M, 3.3M, …, 30M mapped reads;
3. We computed the expression levels for all 20,051 genes in WormBase190 as reads per kilobase of exonic region per million mapped reads (RPKM), using RSEQTools (*18*). As a gene model, we used the "composite", i.e. the union of exonic nucleotides of all isoforms of a gene.

Fig. S4A reports the density plots at the different sequencing depths. As expected, the low-coverage case shows a higher fraction of non-expressed genes. Interestingly, genes which have a $\log_2$(RPKM+1) greater than 2 seem to be less affected by sequencing depth. Fig. S4B reports the comparison between the density plots at different levels of coverage, suggesting that, with a sequencing depth of ~13M mapped reads, most of the expressed genes are captured. This hypothesis is also supported by Fig. S4C, which reports the number of non-expressed genes (RPKM=0) as a function of sequencing depth. Indeed, after ~13M mapped reads the number of "genes" with zero expression begins to plateau, although there remain small numbers of lowly-expressed transcripts that can only be identified by further increases in depth.

C.2. More Detail on "Expression Dynamics" I: Differential Expression

C.2.a. Determining Over-represented Transcripts at Particular Stages

We identified a set of transcripts that are over-expressed in each of the seven main developmental stages (EE, LE, L1, L2, L3, L4, and YA) relative to other stages (Fig. S12). The stage specific transcripts were defined as those highly expressed in a particular stage (>90%) but lowly expressed in at least 4 other stages (<70%). Promoter sequences (-1kb to 0 upstream of TSS) for each group were retrieved and searched for enriched motifs using the MEME algorithm (*19, 20*). To remove generic motifs that are present in promoters of all transcripts, we scanned and compared the occurrences of these candidate motifs in specific transcripts of all the 7 stages. As an example, MEME identified 24 candidate motifs that were enriched in EE-specific transcripts, 12 of which were over-represented in the promoters of EE- or LE-specific transcripts but not in other stage-specific transcripts or ubiquitous transcripts (Fig. S12).

Transcripts for more than 95% of genes were detected in more than one stage in the RNA-seq timecourse, and almost half the transcripts were detected in every stage (Fig. S11). In contrast, only a small number of genes (~100/stage) showed strong stage-specific expression (Fig. S12), suggesting that differences between stages are due to modulation in expression levels of many genes rather than the presence of discrete stage-specific genes.

C.2.b. Detection of Differential Expression from Tiling Arrays

This section describes tiling array processing related to detection of differentially expressed genes. More details are in a companion paper (*21*).

RNA was isolated from 25 different embryonic and larval cell types and from all cells derived from 5 selected developmental stages to generate a total of 30 tiling array data sets (*22-24*). Additionally, 7 tiling array data sets were generated from RNA extracted from synchronized populations of whole animals at 7 different developmental stages. The *C. elegans* Affymetrix 1.0R tiling array was used for all experiments. Non-redundant Transcriptionally Active Regions (nrTARs) were determined by a machine learning approach (*25, 26*)(Note that nrTARs were defined slightly differently than conventional TARs). nrTARs with $\geq$ 20 nt overlap with WormBase coding exons or exons of integrated transcript models were counted as hits. For quantification of transcript levels for annotated genes, unique tiling array PM probes wholly contained within exons of gene models were selected to generate a probe set for each gene listed in WormBase version WS199 (obtained from ftp://ftp.wormbase.org as a gff3 format file). Tiling array data sets were quantile normalized and probe sets were median polished using RMA (*27-29*). Significantly expressed (< 5% FDR) gene models were determined by comparison to an empirical null model of background expression from intergenic probes for each microarray data set (*30*). The total number of detected genes was calculated from the union of tiling array data sets for cells (30 data sets) stages (7 data sets) and for the combination of cells and stages (37 data sets). As a conservative measure to correct for the accumulation of potential type 1 (false positive) errors, we adjusted the q-value of each detected gene by dividing by the cumulative number of independent samples used for each of these estimates (i.e., 37 for cells and stages, 30 for cells, and 7 for stages). This adjustment applied a similar reasoning as Bonferroni correction of *p*-values by assuming that in the least favorable case, false positives, but not true positives, were independent (*31*). To define genes differentially expressed in cells, tiling array results obtained from specific cell types were compared to corresponding developmentally matched reference data sets obtained from all cells. Similarly, to define genes differentially expressed by stage, the 7 tiling array data sets obtained from staged whole animals were compared to each other. Differentially expressed gene models were estimated with a linear model and moderated t-statistic (*32, 33*). Gene models with a FDR $\leq$ 0.05 and fold change $\geq$ 2 were called significant. Differentially expressed genes detected in cells and/or stages were tabulated from the union of the corresponding comparisons. The estimates were adjusted with a Bonferroni-type correction in which the FDR threshold was divided by the number of comparisons between samples. For differentially expressed genes detected in the 25 cell types, the FDR was corrected by the total number of independent comparisons (total of 25). For stages, the FDR threshold was corrected by the total number of pairwise comparisons between data sets derived from seven stages (total of 21) (see Table S4 footnotes 4, 5 and 6, 7). The fraction of genes differentially expressed was determined by dividing the number of differentially expressed genes for each category by the number of genes detected as expressed in the same category (e.g., 11,229 genes differentially expressed in cells and stages divided by 14,279 genes expressed in cells and stages = 79%).

C.3. More Detail on "Expression Dynamics" II: Global Analysis of the Dynamics of Transcription and Binding

C.3.a. Determining a Non-Redundant List of Transcripts and Directly Analyzing their Expression and Binding

In this section we describe how we derived a high-quality list of non-redundant TSSs for studying expression and binding dynamics in Fig. 2A and 2B. This restrictive list has no transcripts that overlap and for each transcript the closest TSS is farther than 0.5 kb away. To derive the list we started with a list of transcripts obtained from WormBase. For each set of potentially overlapping transcripts at a given locus, we kept the longest one and discarded the rest. Then for each kept transcript, we defined a promoter region as a 1 kb window centered on the TSS. In some cases, promoter regions selected in this manner will overlap with other regulatory regions or transcripts, and cause RNA Pol II signal from potentially unrelated promoter regions to enter the window. To minimize this side effect and to reduce double-counting of signal, we found all TSSs less than 500 bp (i.e. half the window size) apart and picked one from each set. Using this approach, we obtained a final set of 8,428 TSSs (and associated transcripts) used for our analyses. RNA Pol II binding levels were obtained by aggregating ChIP-seq signal over promoters. Expression levels were derived from RNA-seq experiments.

Additionally, to further examine our hypothesis that transcripts in early embryo may be inherited from the parent, we compared expression levels from RNA-seq data to RNA Pol II binding sites identified by ChIP-seq in early embryo. We identified 407 transcripts with a DCPM (depth of coverage per million mapped reads, a measure of expression) cutoff of 2.0, 180 of which also had a Pol II ChIP-seq peak corresponding to the promoter region (44%). Thus, 56% of these highly expressed transcripts do not have a corresponding RNA Pol II peak in the promoter region. This finding lends support to our hypothesis that some transcripts are inherited from the parent and may also explain the correlation we see between expression in earlier stages and binding in later stages. RNA Pol II may be binding to these promoter regions in later stages as the organism manufactures more of its own transcripts and shifts from using those inherited from the parent.

C.3.b. Using PCA for Analyzing Expression Changes across Tissues

This section describes how we performed the principal components analysis (PCA) on the tissue samples in Fig. 2C. Our goal was to analyze the overall variation in the RNA-seq and tiling arrays samples in a consistent fashion and then show how the tiling arrays of specific matched embryo-larval pairs show a similar pattern. The embryonic cells were isolated by Fluorescence-Activated Cell Sorting (FACS) of fluorescently tagged cells that had been extracted from dissociated embryos and cultured for 24 hours to allow further differentiation (*21*). The matched tissues from the L2 stage were isolated by precipitation of PolyA Binding Protein.

First, we used gene-expression values for each of the tiling array samples determined in a way as to maximize compatibility with the RNA-seq DCPM values (see description above in supplement sect. B.3.b) and generated two PCAs: a larger one on all the tiling arrays samples as a whole and a smaller one using just a set of matched tissue samples. The larger PCA is described below in the batch effects section and has the advantage that its axes are most compatible with those for RNA-seq.

The smaller PCA in Fig. 2C was produced using 6 pairs of matched tissue samples from mixed embryo (MxE) and L2, giving rise to a 12 sample x 20,085 gene data matrix on which principal components analysis was performed. A 12x12 matrix of principal components was obtained and the matched tissues plotted along the first two principal components, which comprised 67% of the total variation (50% + 17%). In this plot, the MxE and L2 stages are largely separated from each other along one axis while tissue types within a stage are arranged along the other component shown. For both principal components, transcripts contributing most heavily were associated with the GO categories “larval development,” “nematode larval development,” and “post-embryonic development” (Benjamini-Hochberg corrected p-values ranging from 6.9×10^{-61} to 2.4×10^{-70}; hypergeometric test). MxE and L2 cell type samples were largely separated along one of the axes, consistent with the idea that differences in gene-expression programs at the different developmental phases occur across tissues. The observed separation potentially could also reflect differences in sample preparation but we show in sect. C.3.c below, that this possibility can be largely discounted.

C.3.c. Analysis of Potential Batch Effects using Tissue PCA

To examine the possibility of batch effects in the tiling array dataset we repeated the principal components analysis using 42 different samples from the tiling array data set. Different tissues as well as whole animal experiments are included in this larger dataset. This approach gave rise to a large 42 sample x 20,085 gene data matrix. We then applied PCA to this matrix to reduce dimensionality and to identify axes of variance, generating a 42x42 matrix of principal components. The main component of this PCA was particularly enriched for genes with associated GO terms "nematode larval development,” “larval development,” and “post-embryonic development and growth." (Benjamini-Hochberg corrected p-values ranging from 1.5e-113 to 8.3e-113). The variance of the second principal component is primarily explained by a single gonad sample and hence we are not including this component in our batch effect analysis. Additionally, we compared the overall PCA of all the tiling array experiments to that of the RNA-seq experiments (obtained from the correlation matrix in Fig. 2A). Both PCAs shared similar top components.

We examined the PCA of the 42 different samples for possible batch effects using a combination of the Student’s t-test, Pearson correlation, and Spearman rank correlation. Specifically, we examined the effects of lab (Fig. S13), the dates of first and last hybridization, polyA enrichment and the stage. Since two labs generated tiling microarray data, to examine possible lab effects we identified the component value associated with each sample along the first principal component. For the first principal component, these values were used to create two distributions (one per lab) and the means of these distributions were compared using the Student’s t-test. The p-value obtained from this test was p=.07, indicating that variance due to lab is not a significant source of variance along this component. The above process and test was repeated for polyA enrichment (necessarily larval tissue samples) compared to the total RNA preparations and produced a p-value of p=.0007, indicating polyA enrichment does play a role in the variance described by the first principal component.

For comparisons involving more than two states such as date of first hybridization, date of last hybridization, and stage, both Pearson and Spearman correlations were run to determine whether their effects on the first principal component was significant. In the case of comparisons involving dates, samples were binned by quarter and year. For stages, samples were binned as belonging to embryo, larval, or other. Overall, at the level on the variance described by the first principal component, batch effects related to date of first hybridization, date of last hybridization, and lab were not significant at the .05 level. Thus, for the characteristics tested our batch effect analysis revealed significant effects for only stage and polyA enrichment at the .05 level.

The polyA enrichment effect is potentially an issue in relation to the small-PCA analysis in Fig. 2C. That is, in this figure the MxE RNA samples were isolated from sorted cultured embryo cells and the L2 RNA samples isolated by immunoprecipitation of tissue specifically expressed polyA binding protein from whole animal extracts.

To test whether this was significant we plotted the projection of each of the matched tissues from Fig. 2C onto the first principal component of the 42-sample tissue PCA (not the first component shown in Fig. 2C). We find that projections for the MxE tissues range from -0.0092 for GABA neurons to 0.0898 for pan-neural while the projections for L2 range from -0.1958 for GABA neurons to -0.0495 for body-wall muscle. The one exception is L2 pan-neural which has a projection of 0.1018. Since we are now referring to the samples in the 42-sample tissue PCA, we can add the reference datasets into the analysis: the projections of the whole-animal references prepared without any polyA-enrichment are -0.027 for MxE and -0.1564 for L2. On the first component, which represents most of the variation in the data, the L2 tissues clearly segregate with the L2 reference, even though the latter did not have the polyA-enrichment and all are separate from the MxE tissues and reference (the one exception, of course, being the L2 pan-neural sample). One gets a similar clustering when looking at projections on additional components, though the first-component projection provides the simplest and most concise summary.

C.4. More Detail on "Alternative Splicing"

We developed a number of approaches to analyze alternative splicing. A first approach looks at differential splice junction usage. Next, we have two alternative methods to resolve the expression level of individual isoforms for the same gene by distributing RNA-seq reads among a set of alternative transcripts in a probabilistic manner. One method uses expectation-maximization (EM), and the other, a Bayesian approach with Gibbs sampling. We compared relative and absolute expression of alternative transcripts, as identified by either method, between paired samples and across the entire time course of development.

C.4.a. Differential Splice Junction Usage

We created a non-redundant set of all splice junctions, noting the number of reads which confirmed that intron in each stage. We converted that number into reads per million (RPM) by multiplying 1,000,000, and dividing by the number of aligned reads in that stage. We further tracked the depth of coverage per million reads (DCPM) of a transcript that contained that splice junction. To identify alternative isoforms, we sorted the splice junctions by strand and by intron start position (donor), looked at the coordinates of one intron, and asked if the next intron in the list had a start which was equal to the preceding one (same donor, different acceptor), or if it came after the previous but before the next acceptor, etc. For the splice junctions with alternative forms, we looked at how the ratio of the RPM of the two (or more) forms varies over the stages. In this way the control was internal, the path through the region must use one of the splice junctions, and a change in the ratio means differential splice junction usage. To identify examples we performed pairwise comparisons by stage (e.g. comparing the early embryo to the young adult) looking for intron pairs where the transcripts involved both had a DCPM of at least one, where one splice junction in the pair was used at least 5 times more frequently in one stage and less frequently in the other stage, and where at least one splice junction in each pair had an RPM of at least 2 (corresponding to ~5 or more reads for the stages with 25M reads aligned). After identifying candidates in this way, we viewed the change in splice junction usage across stages using a normalized read count for each intron in each stage, calculated by dividing the RPM for that intron by the DCPM of a transcript containing that intron.

C.4.b. IQSeq Analysis

The first method, which we call "IQSeq", uses an expectation-maximization (EM) algorithm to resolve the maximum likelihood (MLE) expression level of individual isoforms. An implementation of this method can be found online (*34*).

C.4.b.i. IQSeq Formalism

IQSeq models RNA-seq as a partial sampling process. Let $I = \{I_1, ..., I_K\}$ be all the possible isoforms for a given gene, with relative abundances $\Theta = (\theta_1, ..., \theta_K)^T$, where $\sum_{k=1}^{K} \theta_k = 1$. We assume that there are M different partial sampling methods (sequencing techniques with difference characteristics, e.g. long/medium/short, single/paired end): $Samp_1, ..., Samp_M$, and let S denote all the samples (reads): $S = \{\text{s from } Samp_m | m = 1, ..., M\}$. We also define $\delta_{s,k}$ as *Ind*(partial sample (read) s is compatible with I_k), where *Ind* is the indicator function. There are in total $N = \sum_{m=1}^{M} N_m$ samples, where N_m is the total number of partial samples from $Samp_m$.

Here we assume a two-step sampling process: First, a sampling method *Samp*$_m$ chooses an isoform instance I_k according to Θ. Second, the sampling method generates a partial sample s according to a local partial sample generation model (the read generation function)

$$G_{s,k}^{(m)} = Pr(\text{generating } s | I_k, Samp_m).$$

Given *I* and *S*, IQSeq then estimate Θ such that

$$\hat{\Theta} = argmax_{\Theta} \log(Pr(S|\Theta)).$$

We derive

$$\hat{\Theta} = argmax_{\Theta} \sum_{m=1}^{M} \sum_{s=s_{m,*}} \log \sum_{k=1}^{K} \delta_{s,k} \theta_k G_{s,k}^{(m)}.$$

This problem can then be solved using EM algorithm by introducing a hidden variable

$$Z_{s,k} = Ind(s \text{ is from } I_k).$$

We denote the estimation for Θ in the nth step as $\Theta^{(n)}$, and further define

$$\zeta_{s,k}^{(n)} = \mathbf{E}_{Z|S,\Theta^{(n)}} [Z_{s,k}],$$

which is the expectation of $Z_{s,k}$ given $\Theta^{(n)}$ and the reads S.

We have

$$\zeta_{s,k}^{(n)} = \frac{\delta_{s,k} \theta_k^{(n)} G_{s,k}^{(m)}}{\sum_{k'=1}^{K} \delta_{s,k'} \theta_{k'}^{(n)} G_{s,k'}^{(m)}}.$$

By performing an E step that computes

$$Q^{(n)}(\Theta) = \mathbf{E}_{Z|S,\Theta^{(n)}} [\log(Pr(Z,S|\Theta))],$$

and an M step that maximizes $Q^{(n)}(\Theta)$with constraint $\sum_{k=1}^{K} \theta_k = 1$. We have

$$\theta_k^{(n+1)} = \frac{\sum_{m=1}^{M} \sum_{s=s_{m,*}} \zeta_{s,k}^{(n)}}{N}.$$

C.4.b.ii. Detection of Differential Expression During Development with IQSeq

We applied IQSeq to RNA-seq data of 7 developmental stages (EE, LE, L1, L2, L3, L4, YA) and derived both the relative and absolute RPKMs for all transcripts. Isoform composition for gene *i* in stage *S* is represented by a vector *θ*(*i,S,k*) where the k_{th} component is the relative abundance of isoform *k* in relation to the other isoforms. Between two stages *R* and *S* for a given gene *I,* the difference in abundance vectors gives a measure of the change in isoform usage for a gene. This is represented as $D(i,R,S) = \sum_{/k} ((\theta(i,R,k) - \theta(i,S,k))_2)/k$. The difference *D* is a fractional number between 0 and 1; scores close to 1 indicate dramatic differences in the relative composition of different isoforms of the gene. The histogram in Fig. 1B and Fig. S14 plots the distribution of *D* values for all genes *i.* Overall, the histogram shows that most genes have minor differences in their isoforms, but a small fraction (280 genes out of 12,875 per pairwise comparison, and 1,324 genes show at least a switch in all 21 comparisons) have major and minor isoform switching between stages (major and minor isoform showing at least 5.7 fold difference in expression abundance). We computed similar quantities for absolute differences. Genes are then classified based on their scores in these two statistics in pairwise comparisons, revealing the subsets which show only dramatic isoform composition change, only dramatic absolute expression level change, neither, or both. Further analysis on these subsets may reveal key gene players or pathways in dictating nematode development.

C.4.c. Deepseq9 Analysis

The second method, which we call "deepseq9", uses a Bayesian approach to estimate the relative expression of alternative transcripts for the same gene. An implementation of the algorithm, including documented source code, is available at SourceForge (*35*). Deepseq9 was developed by B. Carpenter (Statistics Dept., Columbia University) and M. Morris (CGSB, NYU).

C.4.c.i. Computing Transcript-level Expression using a Joint Model of Read Alignment and Expression

Given a data set of sequence reads, our goal is to estimate the expression of each alternative transcript for a gene based on the abundance of reads which map to sequences contained within each isoform. The method effectively distributes all of the observed reads among the possible isoforms using a probabilistic logic. Briefly, expression is inferred from the following data:

$K \in N^+$ (the number of variant isoforms), $N \in N^+$ (the number of reads), and $y_1,...,y_N$ (the reads).We assume two model hyperparameters: φ (the expected variation from the reference sequence), and $\alpha_1,...,\alpha_K \in R^+$ (the prior read count per sequence plus one (to avoid zero division errors)). The general-purpose parameter vector φ reflects deviation of the sample sequence from the reference sequence for the given read distribution due to factors such as SNPs, amplification errors during sample preparation, and the sequencing platform's error profile. We infer two model parameters: $t_1,...,t_N \in 1{:}K$ (the mapping of read to splice variant), and $\theta_1,..., \theta_K \in [0,1]$ such that $\sum_{k=1}^{K} \theta_k = 1$. (Note, θ in Deepseq9 and IQSeq are equivalently defined.)

- $\theta \sim \text{Dirichlet}(\alpha)$
- $t_n \sim \text{Discrete}(\theta)$ for $n \in 1{:}N$
- $y_n \sim \text{Channel}(t_n,\varphi)$ for $n \in 1{:}N$

To estimate expression levels, we must calculate the posterior probability of reads mapping to all possible alternative transcripts. The model uses Gibbs sampling to draw samples from the full posterior distribution $p(\theta,t|y,\alpha, \varphi)$ computed over read mappings t_n and read expression levels θ given the reads y, resulting in a discrete sampling of the mappings t_n onto all annotated isoform variants based on the parameter θ (effectively a beta-binomial model of expression level). The read channel model assigns the probability of a given read y_n being observed, given that it arose from the splice variant t_n under the model parameterized by φ.

C.4.c.ii. Alternative Splicing in the Aggregate Integrated Transcript Set

The analysis was initiated using pre-computed exon-level coverage for the annotated aggregate integrated transcript models, expressed in DCPM, and a count of mappable reads for each exon (DCPM_bases), as determined from initial mapping of the RNA-seq data to the *C. elegans* genome (WS190) as described above (see sections above C.1.a and C.1.b). For each exon, we generated a set of putative alignments to all parent transcripts, and then used our Bayesian model to jointly compute the read assignment and transcript-level expression. The alignment score is the probability of the read given the exon, which is proportional to the exon length (counting only mappable bases): $P(\text{read}|\text{exon}) = \log_2(\text{ExonLength}/\text{TranscriptLength})$. We multiplied DCPM by 1000 to obtain pseudo-reads that align to the exon and then generated mappings between each pseudo-read and each possible parent transcript. The average number of mappings to distinct transcripts per read was 3.1 (i.e., on average, reads for each exon could map to one or more of three alternative transcripts). For the deepseq9 expression program, the Gibbs sampler was run for 1000 epochs, with a burn-in parameter of 500 (i.e., the first 500 iterations were discarded to allow the model to reach a stationary distribution); thereafter, we took one sample every 10 epochs (thinning of samples in this way reduces the effect of auto-correlation on samples and produces better variance estimates with fewer samples). Expression was computed as the average number of reads per transcript across all the samples. We compared our estimates with extrapolated transcript DCPM counts from the initial mapping described above, and found good overall correlation between the two approaches (median R^2=0.82 across the 15 samples).

C.4.c.iii. Clustering Expression by Developmental Stage using Self-Organizing Maps (SOMs)

We combined the transcript-level expression calculated by deepseq9 for all aggregate integrated transcripts across the 15 stages into a single data table. To identify alternative transcripts which show a relative change in expression (i.e., transcript A > transcript B in stage 1; transcript A < transcript B in stage 2), we applied filtering criteria requiring that: (a) transcripts differ by at least 30% in opposite directions in at least two stages, and (b) the more highly expressed transcript has at least 5 pseudo-reads (corresponding to a DCPM of 0.005). (We note that ~800 transcript pairs which passed these filters displayed borderline expression levels due to the low minimum read threshold, thus resulting in lower confidence estimates of differential expression.) The set of transcripts that passed these filters (15,064 transcripts for 3,428 genes) was run through an SOM clustering algorithm (R 2.11 - library(class), function "SOM") that generated 48 clusters (Fig. S15).

C.4.c.iv. Identification of Alternative Transcripts with Different Developmental Profiles

We found that 43% of all genes subjected to clustering showed alternative overlapping transcripts which fell into two or more different SOM clusters (corresponding to 7,203 transcripts for 1,475 genes) (Fig. S16). From a total of 4,846 pairs of clusters containing alternative overlapping transcripts for the same gene, we further examined 2,788 cases (involving 5,443 transcripts for 1,320 genes) in which precisely one isoform fell into a distinct cluster from other isoforms for the same gene. Among these we were able to discern several distinct classes of alterations in features at the 5' end, within the CDS, or at the 3' end of transcripts (Fig. S17 and (*36*)).

Individual examples from these different classes are shown in Fig. S18. We observed that while most cluster pairs shared fewer than 4 genes, those pairs with the largest number (proportion) of genes in common also tended to show similar developmental profiles. Thus, for follow-up of individual genes, examples from cluster pairs with fewer genes in common are more likely to reveal alternative transcripts with more obviously divergent developmental expression profiles.

C.4.d. Validation of Inferred Transcript Expression from Deepseq9 and SOMs

We used two methods to empirically test support for the differential expression of alternative transcripts during development inferred by deepseq9 and SOM clustering: (1) validation using qRT-PCR on specific examples presented in Fig. S18, and (2) comparison of the DCPM values from staged RNA-seq data estimated by deepseq9 using exonic reads (upon which we based our clustering) with counts from the same datasets of reads that span exon junctions, for the set of isoforms represented in Fig. S17.

C.4.d.i. Overview of Validation using qRT-PCR

Isoform-specific qRT-PCR was conducted using staged *C. elegans* RNA samples for two genes, F26B1.2 and C25H3.7 (Fig. S18). These were selected for validation tests based on their distinctive cluster patterns, the presence of diagnostic splice variants, and high expression counts (the third example from Fig. S18 was not tested due to its lack of diagnostic splice junctions). Primers were designed to amplify specific isoforms or isoform groups for each gene (illustrated in Fig. S18): F26B1.2 transcripts 8 vs. 9 (F26B1.2-T8 vs. F26B1.2-T9); and C25H3.7 transcript 3 (C25H3.7-T3) vs. the transcript group 1/2/4 (C25H3.7-T1, C25H3.7-T2, C25H3.7-T4). Transcript expression levels were calculated in terms of the fold-change formula FC = $2^{-\Delta Ct}$ using *act-1* as an endogenous control (*37*), which is sensitive to fold-changes of 2 or greater.

To compare the two methods, we evaluated whether the direction of change in relative expression levels measured by qRT-PCR matched that estimated by deepseq9 using RNA-seq DCPM counts. In both cases, we obtained consistent results. For F26B1.2, qRT-PCR showed that T8 is expressed 25-fold higher and 2-fold higher than T9 in L2 and L4, respectively; thus, the difference in expression between T8 and T9 decreases from L2 to L4 according to both methods. Similarly, for C25H3.7, T1/T2/T4 is expressed 35.5-fold higher and 191-fold higher than T3 in Embryo and Young Adult, respectively; thus, the fold difference in expression between T1/T2/T4 and T3 is greater in Young Adult than in Embryo according to both methods. Overall, the qRT-PCR data for genes F26B1.2 and C25H3.7 show consistent trends that qualitatively support changes in relative isoform expression estimated from the RNA-seq data using deepseq9.

C.4.d.ii. Materials and Methods for Validation using qRT-PCR

To collect staged samples, N2 worms were bleached in 20% alkaline hypochlorite solution for three minutes and washed four times with M9. Embryos were either collected immediately for RNA extraction or rotated in M9 overnight to hatch and arrest at the L1 stage. L1s were either collected immediately for RNA extraction or plated on OP50-1 seeded 150cm NGM plates to a capacity of 20,000 worms per plate in order to develop to the various larval stages. Plates were incubated for 22 hours at 15°C, 20°C, and 25°C to collect L2, L3 and L4 worms, respectively. For the young adult stage, plates were continually kept at 20°C for two days after plating L1 worms. Each stage following L1 was characterized by size and morphological markers.

RNA from the various stages was extracted using the RNAeasy Kit (Qiagen). cDNA was prepared using SuperScriptIII Reverse Transcriptase (Invitrogen) and a poly(dT) primer. For gene expression analysis one microliter of cDNA was used in a SYBR Green qRT-PCR reaction (LightCycler FastStart DNA MasterPlus SYBR Green Kit, Roche). qPCR was performed on a LightCycler 480 (Roche) with primer annealing at 58°C and florescence capture after extension. Crossing-threshold (Ct) values were calculated using the LightCycler 480 Sofware (Roche).

Isoform-specific PCR primers with the following sequences were designed to produce products that span diagnostic introns that are not shared between the relevant pairs of isoforms, as shown in Fig. S18:

C25H3.7.T3-F: 5'-TCGGTTTCTGGATCGAAGAT-3'; C25H3.7.T3-R: 5'-TCCTTTGGCAAGGTAGTTGG-3'; C25H3.7.T124-F 5'-CGTCAATCTCCACGAGGACT-3'; C25H3.7.T124-R 5'-GCATTGTTCACAGTTTTGTCG-3'; F26B1.2.T8-F 5'-CGAGAGCACGATAATGACGA-3'; F26B1.2.T8-R 5'-TTTTTTTTTTTTGAGAACAGTCTTCTC-3';F26B1.2.T9-F 5'-CAAAGTGGGAGCCGCTATTA-3'; F26B1.2.T9-R 5'- AGCATGCGCACTTCACAC-3';.

Primer sequences for the *act-1* control were:

act-1-F 5’-GCTGGACGTGATCTTACTGATTACC-3’;*act-1*-R 5’-GTAGCAGAGCTTCTCCTTGATGTC-3’.

C.4.d.iii. Validation using Reads Spanning Exon Junctions

The estimated transcript-level DCPM values from deepseq9, upon which we based our clustering (Fig. S15, S16), were computed using only those reads that map fully within a single exon. Therefore, comparisons of exon junction-spanning reads from staged RNA-seq data with the deepseq9 DCPM values provide independent evaluation of the results from deepseq9. We selected all introns that differ between pairs of isoforms based on their presence or absence in each model, and selected the reads spanning the flanking exons, which we call “discriminative reads”. For each transcript, we then compiled a developmental expression profile based solely on the total counts of the discriminative reads for each stage.

As a result of the clustering described above, we identified 2,788 cases involving 5,443 transcripts for 1,320 genes where precisely one isoform fell into one cluster, and one or more isoforms fell into a different cluster (Fig. S17). In 2,408 of these cases, involving 7,208 individual transcripts, an alternative splicing event was involved (as opposed to transcripts differing only by extensions of terminal exons). Of these, a total of 2,002 cases involving 6,149 transcripts for 1,009 genes had sufficient counts of discriminative reads that we could use them for our comparisons, resulting in 5,733 possible pair-wise comparisons between alternative isoforms.

We asked if the expression profiles for the discriminative reads showed the same level of differential stage-specific switching as the DCPM counts, which we previously defined as a 30% difference in isoform abundance in opposite directions in at least two stages. Among the 5,733 transcript pairs that we could compare (all of which showed stage-specific switching in the DCPM profiles), 50% (2,889) showed stage-specific differential switching in the discriminative read profiles, with an average Pearson correlation of 0.76 between the two sets of profiles. Thus, while the discriminative reads show less variation overall than the DCPM counts, the profiles that do show variation correlate well with the DCPM data across the 15 stages used in the clustering. This is a very demanding test because the total counts for discriminative reads are lower and thus noisier than the DCPM counts, show less variation overall as noted above, and are not length-normalized as are the DCPM data. More specifically, of the cases where we found informative reads spanning the splice junctions only 52% of them had sufficient depth of coverage on the splice (>5 reads) to allow accurate quantitation.

C.5. More Detail on “Pseudogenes”

C.5.a. Pseudogene Assignment

Pseudogenes are usually identified by the rapid accumulation of mutations such as premature stop codons. They are created from protein coding genes either by duplication followed by disablement or from the integration of reverse transcription of processed transcripts into the DNA (*38*). By definition a pseudogene is not functional in the conventional sense of its protein-coding parent. However, this does not preclude the "dead" pseuodgene sequence from acquiring new functions either as a transcribed RNA or even as a translated peptide. There have been a number of reports of and speculation on such apparently "revived" pseudogenes (*38-42*).

In order to identify a list of possible *C. elegans* pseudogenes, we looked at a number of features including amino acid sequence identity, how much the pseudogene covers the parental gene, and modifications such as insertions, deletions, and frame-shifts. This analysis was performed both by using the automated pipeline PseudoPipe (*43*) and by hand-annotating the *C. elegans* genome with the help of data available in the WormBase database. Comparing the coordinates from the 2,343 candidate pseudogenes identified by PseudoPipe and 1,541 identified by WormBase, there were 1,025 pseudogenes which had a nucleotide overlap of at least 50 bp between the candidates in each data set. The remaining sequences were reviewed manually, and it was determined that 173 pseudogenes from PseudoPipe and 95 pseudogenes from WormBase should also be included in the list, for a final total of 1,293 (Fig. S19). The remaining sequences either overlapped with annotated genes, were too small and fragmented to be considered a pseudogene, or should have been curated as part of a functional gene. We also established the probable source (parent) gene for 1,198 pseudogenes.

C.5.b. Pseudogene Transcription

We investigated the 1,198 pseudogenes with identified parent genes for evidence of transcription based on the RNA-seq data. We found 323 of them to be abundantly expressed using the RNA-seq read mapping procedure described in (*14*). In this method, all reads were aligned (using MAQ and cross_match) to the genome, splice junctions spliced leaders, and polyA libraries. The best match was then chosen with only a minor bias for a genome match first - reads with equal matches to the genome and other databases were placed against the genome. The DCPM was calculated from these mapped reads. Pseudogenes were determined to be transcribed if they had a DCPM value of >0.04 in at least one sample. This threshold is 100-fold higher than the minimum DCPM value in this set.

In order to address the possibility that the reads were derived from the parent gene and not the pseudogene, we classified the pseudogenes into three subcategories. The first includes pseudogenes with expression levels at least two-fold higher than the parent gene. The second subclass contains pseudogenes for which the expression patterns of the pseudogene and parent are discordant across samples (see Fig. 1D for an example). Both of these cases indicate independent transcription of pseudogene and parent, arguing against mapping artifacts. The last subclass includes instances where the expression pattern of the pseudogene is concordant with the parent gene across multiple samples, which by itself would not exclude mapping artifacts. Altogether 191 of the 323 candidates fell into the first two subclasses (87 and 104, respectively) and are thus likely transcribed independently from their parents.

C.6. More Detail on "ncRNAs"

C.6.a. Identification of Canonical miRNAs and Mirtrons

Canonical miRNAs are produced by sequential cleavage of inverted repeat transcripts by the Drosha and Dicer RNAse III enzymes. We annotated novel canonical miRNAs using the miRDeep algorithm (*44*), and for confident annotation, required that the cloning of miRNA and star reads mapped to a precursor hairpin with 3' overhangs at both ends of the inferred small RNA duplex. A subset of loci was confirmed to be dependent on the Argonaute encoded by *alg-1* (*45*). In total, 24 confident novel miRNAs were deposited in the miRBase database.

For mirtrons, we built an SVM model based on features of the 14 initially reported *D. melanogaster* mirtrons (*46, 47*) and ran this on the *C. elegans* genome as an independent test of its performance (*48*). Three of the four known nematode mirtrons (*mir-1018*, *mir-62* and *mir-1020*) ranked within the first 27 candidates genomewide; the fourth (*mir-1019*) presents a highly atypical 2:5 hairpin overhang and scored much lower (554th). We validated high-scoring predictions using publicly available small RNA data (*45, 49-57*), yielding 12 novel mirtrons that produced at least 5 small RNA reads with a dominant 5' end and extending to the intron terminus; 10 of these also generated star reads with appropriate duplex overhangs. *NM_075944_in2* and *NM_071513_in8* did not have star reads, but the recovery of >40 reads from both loci with precise 5' ends provided strong evidence of specific miRNA production. Several other loci with candidate evidence (i.e. <5 intron-terminal reads) were noted, which may reach strong confidence with additional sequence data. We also reclassified the previously annotated *mir-2220* as a mirtron and recognized *NM_075943_in1* to produce a mirtron from an unannotated splice site, for a total of 18 confident mirtrons in *C. elegans* at present. Several additional high-scoring predictions yielded <5 intron terminal reads and were classified as candidates. Full analysis of mirtrons in *C. elegans* is available at (*58*).

C.6.b. Predicting Novel ncRNA Candidates

C.6.b.i. Known ncRNAs

The genome produces a variety of transcripts that do not code for proteins and function directly as RNA (non-coding or ncRNAs). Altogether at the start of the project there were 1061 known ncRNAs in *C. elegans* (Table S5). These include small and medium size RNAs (e.g. miRNA, snRNA, snoRNA, etc), and also long RNAs (e.g., rRNAs etc.) involved in mRNA translation and splicing.

C.6.b.ii. Building the 7k-set of Novel ncRNA Candidates

First, using small RNA sequencing data alone, we defined 102 additional candidate canonical miRNAs (*45*), of which 42 have evidence of complementary strand sequence (providing confirmation of the miRNA) and of which 20 were recently incorporated into miRBase (*59*). We further tried to separate known ncRNAs from other genomic elements such as CDSs and UTRs, but found that they cannot be separated completely by any single genomic feature (e.g. conservation, small RNA sequencing data) (Fig. S21A, Left). Although discrimination improved when pairs of features were examined, it was still incomplete (Fig. S21A, Right). To address this incompleteness, we used a machine learning method to integrate nine genomic features to identify ncRNAs and predicted 7,237 novel ncRNA candidates (7K-set)(*60*). An example of a candidate ncRNA is shown in Fig. S21B. We found that many novel ncRNA candidates were expressed in embryos, which suggests the development functions of these candidates. Next, we also clustered novel ncRNA candidates with coding transcripts using the total RNA tiling array data. From GO analysis (Table S9), we found that some novel ncRNA candidates were clustered with coding genes that are DNA binding proteins and transcription factors.

C.6.b.iii. Building the 21k-set of Novel ncRNA Candidates

In addition to the 7K-set (*60*), we describe below how we constructed the 21K-set of ncRNAs. The construction of the 21K-set follows similar principles as those described above for the 7K set. However, it does not include DNA conservation and RNA secondary structure information.

The tiling array signals were segmented into TARs (Transcriptionally Active Regions) using the maxgap/minrun algorithm (*12, 13*). Briefly, a contiguous sequence of probes exceeding a signal threshold (selected as described below) was connected to form a TAR. To account for noise, a total of 30 bp (about 1 probe) were allowed to fall below this threshold within a single TAR. Finally, TARs shorter than 100 bp (the total length of 4 probes) were discarded. The signal threshold was optimally selected according to the criteria of attaining an FPR of 0.05 when compared to a high confidence subset of the annotation. Details are provided in (*11*).

In total, 95,069 TARs (37,026,882 nt in total) were collected from the union of 41 tiling array experiments (Table S3), of which the minimum length is 100 nt. 1,331 overlap with known ncRNA, and 22,487 include transcribed regions that are not overlapped with any annotated (confirmed or predicted) exons or known ncRNA. The reads from sequencing data from small RNA and polyA-selected RNA were also averaged for each tiling array TAR. Subsequently, different types of expression values were combined to classify each TAR as ncRNA, CDS, or UTR, using machine learning methods. Known ncRNAs, CDSs, and UTRs were selected as a gold-standard set for machine learning (Tables S6-8). Before classification, the 95,069 TARs were fragmented into 448,746 small windows (using sliding windows of 150 nt with a 75 nt step size) (Fig. S20). Because of the sample preparation method, the tiling array TAR cannot inform as to which strand the transcript came from.

Although lacking conservation and secondary structure information, the accuracy of the classification model for the gold-standard set in terms of AUC (area under the ROC curve) is still as high as 94.2% for ncRNA prediction from TARs (Table S7). When applying the classification model to the 49,648 novel transcribed windows (from 22,487 TARs), 45,913 were found most likely to be ncRNA, 3,294 were most likely to be UTR, and 441 were most likely to be CDS (Table S8). These 45,913 "windows" originated from 21,521 TARs out of the original set of 95,069 TARs. This gave rise to the 21,521 predicted ncRNA TARs in the 21K-set. Subsequently, 1,259 of the predictions in this set were found to overlap the predicted ncRNAs in the 7K-set. The genome locations of the 21K-set are available at (*36*). Note, the prediction accuracy of the 21K-set is not as high as the 7K-set, and many of them could come from UTRs or unprocessed introns.

D. More Detail on "Regulatory Sites and Interactions"

D.1. More Detail on "TF-Binding Sites, Motifs, and Targets"

D.1.a. Overview of the Experiments

To date, large-scale projects aimed at mapping TF binding sites have been performed either in cell culture or in single-celled organisms, and have failed to link the identified regulatory elements to developmental events. We investigated binding sites within the whole animal using high-throughput sequencing ChIP (ChIP-seq) to map 23 GFP-tagged fusion proteins. Generally, the factors were mapped at the developmental stages during which they have their highest expression levels, as deduced using Green Fluorescent Protein (GFP) fusion proteins. At least two independent ChIP-seq experiments were performed for each factor.

D.1.b. Scoring: Broad Regions and Narrow Summit Peaks

The binding sites (broad regions) of each factor were scored using PeakSeq with a q-value threshold of 0.001 (*61*). Initially, we scored the peaks using the pooled reads from two replicas. This gave to an initial set of broad binding regions. We used these for determining targets and to define HOT regions (see details in D.1.e. and D.2, respectively).

Next, we progressively filtered this set to refine our peak calls. First, we required regions to overlap between the regions between replicates, creating a subset of broad regions (the overlapped set of broad binding regions). We only kept those regions reproduced in both replicates and any regions less than 50 nt were removed (*62*). We refined region summits, using PeakRanger to identify the multiple summits inside each broad region found by PeakSeq (Fig. S24). PeakRanger takes the broad regions and uses second-derivative information to find local summits. In PeakRanger raw read signals are first enhanced to reduce the potential of producing false positives and then traversed using a summit detection algorithm.

We extended each of these summits with a 100 bp flanking each side (if the initial region was smaller than 200 bp we just kept the small region size). The narrow binding regions (maximum 200 bp) around the summit found by PeakRanger were defined as "narrow summit peaks". The rationale for using 200 bp as the width is that this is the median shear size used in the ChIP-seq experiments. The 200 bp peak width was further validated by the conservation calculation in sect. F.3.c.

The pipeline yielded a final set of summit peaks that was used for genome accounting, conservation analysis and TF site predictions. The numbers of total mapped reads and the number of narrow peaks for 23 factors are shown in Table S10.

For a comparison, we also analyzed the peak centers using the SPP peak calling algorithm (*63*) and used some of the SPP summits for the motif finding. A strong rationale for keeping the scoring compatible with PeakSeq and SPP is that these two programs were chosen as the "peak callers" for the ENCODE Human project, based on extensive peak caller comparisons. We wanted the scoring in worm modENCODE to be compatible with that in ENCODE, which enabled the comparisons between worm and human in sect. G.

D.1.c. Binding-Site Validation

Control experiments using antibodies directed against native proteins demonstrate that tagged protein binding sites correlate strongly with those from native protein. Also, TF binding sites identified through ChIP-seq have been verified through an independent method, ChIP-qPCR (*62, 64*).

We performed a series of analyses to examine the quality of our ChIP-seq experiments. Much of these are discussed in detail in (*64*) and summarized here. First, we selected several factors (for which primary antibodies are available) to compare our transcription factor (TF) tagging strategy for ChIP to native protein ChIP. We found that: (1) GFP-tagged AMA-1 has the same binding pattern as does native AMA-1 (the correlation coefficient between samples is 0.934), (2) the binding sites of GFP-tagged PHA-4 from embryos and starved L1s are verified by comparing our list of genes to the list of known pharynx developmental genes (90/238, P<1.7e-13), and (3) the binding sites of GFP-tagged HLH-1 were validated by comparing our result to an unpublished data set of binding sites for endogenous HLH-1. Overall, these analyses (to date) are consistent with the conclusion that the tagged factor has binding and regulatory properties similar to those of the native proteins, and that differences between the tagged factor and native protein ChIPs are well within the expected levels of variation which are commonly observed between replicate ChIP samples using the native protein. Second, many PHA-4 binding sites from embryos and starved L1s identified by ChIP-seq were verified through an independent method: ChIP-qPCR (76% of the embryonic sites and 74% of starved L1 sites with two-fold or higher enrichments).

Finally, we calculated the functional enrichments of protein-coding genes targeted by each of the 23 factors. Many Gene Ontology (GO) terms related to developmental processes were enriched for the list of genes bound by many factors in this study, suggesting the general roles of these factors during C. elegans developmental processes. More importantly, for factors with known functional roles we identified specific enrichment of GO terms that match these functional roles (*62*). In conclusion, these analyses demonstrate the high quality of our ChIP-seq experiments.

D.1.d. Identification of TFBS-associated Sequence Motifs

A major characteristic of most TFs are their sequence recognition motifs. These motifs are typically short, inexact sequences ranging in size from 8 to 12 bp (*65*). We developed a technique to identify high-likelihood cis-regulatory motifs from the modENCODE ChIP-seq TF binding data sets. We combined information from both PeakSeq (*61*) and SPP (*63*) with information from the six-way nematode alignment (see Conservation section, main text). For these calculations we excluded the HOT regions (described below and in the main text). We weighted sequences under peaks for each TF by their degree of evolutionary constraint and distance from the peak center. To discover motifs, weighted sequences were processed with the MEME sequence-pattern discovery algorithm (*19*, *20*) (along with background sequence generated by a fourth-order Markov model from peak flanking regions), applying a p-value cut-off of 0.05. Although we used evolutionary constraint to identify putative TF motifs, the presence or absenceof motifs was not used during the analysis of evolutionary constraint under TF binding sites. Each motif predicted by MEME was tested for specificity by measuring the frequency of the motif occurrences in peak regions relative to random upstream sequences and peaks from other TF data sets (Fig. S35C). We also performed localization tests for each motif relative to point binding positions (Fig. S35B). The initial pattern discovery algorithms identified statistically enriched motifs for 21 of the 23 TFBS profiles, but motifs for only 8 TFs remained after specificity testing (Fig. S35A). Of the three TFs with previously described putative binding site motifs, we recovered the previously described motif for HLH-1 and PHA-4, but failed to recover the published motif for SKN-1.

D.1.e. Identification of Target Coding and Non-coding Genes

The details of data sets for 23 factors (22 TFs and one dosage compensation factor) are listed in Table S10. We used the middle point of the binding region to calculate the distance to the TSS of genes and determine the targeted genes for each TF. We used a simple approach: TF binding peaks within 500 bp upstream or 300 bp downstream of a gene's TSS were assigned to that gene. This is a fairly conservative approach; it is possible to take significantly larger values for the upstream threshold without greatly affecting the results. This is borne out by analysis of the aggregation plots in Fig. S22C. Aggregation plots were drawn with ACT (*66*).

We extracted *C. elegans* gene annotations from WormBase. Although the TSSs for the majority of *C. elegans* miRNAs have not been mapped, it has been shown that DNA regions upstream of the pre-miRNA are sufficient to initiate the transcription of miRNAs (*67*). Therefore we identified the target miRNAs by examining the existence of TF binding peaks around the start position of pre-miRNA transcripts.

In comparison to coding genes, binding sites assigned to known ncRNAs are even closer to the 5′ end of the transcripts. Consequently, binding sites could be readily assigned to specific protein-coding or known-ncRNA genes, based on their proximity to the TSS. Most binding sites were assigned to annotated loci, but a subset remained unassigned for each factor. Although most factors bind sites near both protein coding and known ncRNA genes, GEI-11 binds mainly to ncRNAs (Fig. S22). We also examined whether any TF binding sites were adjacent to our previously undescribed novel predicted ncRNAs (intergenic ncRNAs from the 7k-set, see above). Approximately ~59% are potential targets of the 22 TFs examined, significantly more than would be expected by chance ($P < 0.001$, estimated by the z-score, assuming a normal distribution). This provides additional evidence for their activity.

We also compared the targets shared by all TF pairs. Pairwise correlation analysis of target genes revealed that factors with related functions often show substantial overlap in the target genes to which they bind (Fig. S23A), such as MAB-5, LIN-39 and EGL-5. They are more strongly correlated with each other in terms of targets than with the other four HOX genes analyzed, which have more diverse developmental roles. In contrast, factors binding at pairwise correlation of miRNA targets show that the factors bound to them tend to cluster together more by stage than by factor type (Fig. S23B). For example, one group of 4 different TFs analyzed in embryos target similar miRNAs, whereas a different group of six disparate TFs analyzed at L3 target another set of miRNAs. Integrated regulation by multiple TFs at a given developmental stage may have to do with the fact that the expression of miRNAs tends to show strong stage-specific enrichment. The large fraction of the genome associated with sites and the high number of genes targeted from the relatively small set of TFs we analyzed (from >900 candidate TFs in *C. elegans*) suggests that each gene may have sites for many factors.

D.2. More Detail on "Clustered Binding in HOT Regions"

D.2.a. Identification of HOT Regions, with Sensitivity Analysis

Using the 23 factors' ChIP-seq data sets, we determined the number of factors bound at each base in the *C. elegans* genome. Out of the 16,707 genomic regions identified as having significant enrichment in at least one of the ChIP-seq data sets (using the broad peak regions described in D.1, with a *q*-value cutoff of 1e-5), 304 Highly Occupied Target (HOT) regions were significantly enriched in 15 or more factors (*36*). We combined overlapping peak regions across the 23 factors to annotate each of the 16,707 regions based on the maximum number of factors associated at any base point within the called peak. To determine whether this would be expected by chance, we randomly re-assigned peak regions within the 16,707 regions bound by at least 1 factor. Using 1,000 iterations of random re-assignments, no regions associated with 15 or more factors were observed (Fig. S25A).

We next wanted to determine whether HOT regions remained bound by 15 or more factors when peaks were defined more narrowly. Peaks were re-defined as the region 26-, 50-, 100-, 150-, or 200-nt wide centered on a peak summit (identified by PeakRanger as described in D.1). We found that over 80% of HOT regions remained bound by 15 or more factors when peaks were narrowed to 200, 150, or 100-nt wide (Fig. S25B). These results indicate that TF binding in HOT regions occurs within a 100bp window.

We make available through the supplementary website both core 304 HOT regions and alternate lists of HOT regions defined using narrower peaks (*36*).

We used multiple experimental and computational approaches in order to confirm that enrichment for these regions was not simply an artifact of the ChIP-seq procedure. HOT regions were not significantly enriched when IgG antibody was used on transgenic animals or when GFP antibody was used on N2 animals lacking a GFP-tagged TF. These negative controls demonstrate that these regions are not simply a chromatin or GFP-antibody artifact (Fig. S26A). As an additional negative control, we observed that DPY-27, which is known to bind preferentially to the X chromosome (*68*), is almost exclusively enriched at regions (including HOT regions) on the X chromosome and is not enriched at HOT regions on the autosomes (Fig. S26B). As a positive control, we immunoprecipitated endogenous LIN-15B from wild-type animals using anti-LIN-15B antibody, and observed binding peaks in HOT regions similar to those observed using the GFP antibody on *lin-15B*::GFP animals (Fig. S26A).

D.2.b. Expression of Genes Associated with HOT Regions

We used a stringent criterion to associated genes with HOT regions. Genes were associated with peak regions if they were within 1kb upstream or 500nt downstream of the gene’s TSS. For staged populations, gene expression levels for all *C. elegans* WS190 transcripts were measured by DCPM in RNA-seq data as described previously (*14*), and for genes with multiple annotated alternative transcripts, the average expression level of all transcripts was used. We also used two different types of tiling array data sets described in (*21*): tissue-specific embryonic expression measurements (performed by expression of GFP under tissue-specific promoters followed by

FACS sorting), and tissue-enriched measurements (performed by tissue-specific promoter-driven expression of epitope-tagged polyA binding protein followed by purification of RNA bound by the tagged polyA binding protein--the mRNA tagging method described in (*76*)). Tiling array data were analyzed by first computing the PM - MM value for each probe. Experiments were conducted in triplicate and quantile normalization was used to ensure values from the three replicates were comparable. Data from the three replicates were combined using pseudomedian smoothing (*12*) over a window size of 110 bp, and transcript expression levels were calculated as the median signal value for all probes overlapping the transcript's exonic regions by at least 50%. Only the longest isoform was used for genes with multiple transcripts. For inter-sample comparison, we normalized these expression levels by dividing the values by the slide median (i.e. the median of all probes on the array). In the staged population RNA-seq experiments as well as every tissue from both tissue-specific and tissue-enriched tiling array data, HOT genes had significantly higher levels of expression than genes bound by 1-4 factors (all P < 1e-15 based on Kolmogorov-Smirnov test) (Fig. S28).

D.2.c Comparing Targets: Factor-Specific vs. HOT

D.2.c.i. Motif Enrichment and Tissue Specificity

HLH-1 is a muscle-specific TF with a consensus binding motif CAGCTG ((*70*), Fig. S35). Motif enrichment was calculated by simple hexamer frequency counts, and p-values were calculated using the chi-square test. Genes with L1 muscle-enriched expression were obtained from (*69*). To compare all TFs, we additionally made use of L4 intestine-enriched transcripts (*71*) and embryonic tissue-specific tiling arrays described above (*21*). To identify embryonic tissue-specific genes, each embryonic tissue-specific array was first linearly normalized to the embryonic reference array to correct for array-specific scaling effects. Next, for each gene in each tissue, we calculated a z-score for specificity: $z_{tissue} = \frac{x_i - \mu_i}{\sqrt{\frac{1}{N-1}\sum_{j=1,..,i-1,i+1,...,n}(x_j - \mu_i)}}$, where $\mu_i = \frac{\sum_{j=1,..,i-1,i+1,...,n}(x_j)}{N-1}$ and N=11 tissues (including the reference array). Genes with z_{tissue}>2 were deemed "tissue-specific".

For this analysis, we used three well-characterized tissue-types: intestine, hypodermis, and body wall muscle. We identified TF-tissue pairs wherein genes associated with factor-specific peaks for a TF were significantly enriched (above the background set of all genes) for the set of tissue-specific genes (requiring both fold-change greater than 2.5 and P < 1e-5 by Fisher's exact test). TF factor-specific targets were significantly enriched for the tissue-specific expression compared to HOT targets for 13 of the 15 TF-tissue pairs that met this criteria (P < 0.01 by Fisher's exact test), and in 8 cases were still significant at a P<0.0001 cutoff (Fig. S27B). In addition to HLH-1, we considered previously identified binding motifs for ELT-3 (GATAA (*72*)), MDL-1 (CACGTG (*70*)), and PHA-4 (T[AG]TT[TG][AG][CT] (*73*)). For the three additional factors, we observed a drop in motif enrichment between factor-specific targets and HOT regions similar to that observed for HLH-1 (Fig. S27B).

D.2.c.ii. Comparison of Enrichment in Essential Genes

Essential genes were defined as genes having an RNAi phenotype of 100% larval arrest, embryonic lethality, or sterility in a genome-wide screen for RNAi knockdown phenotypes (*74*). Significance was calculated by the chi-square test (Fig. S27C).

D.3. More Detail on "Building a TF Hierarchy"

The network was visualized with Cytoscape (*75*) and analyzed with tYNA (*76*). We examined the difference of TFs at different layers of the hierarchical network we constructed. Specifically, we compared the tissue specificity scores and degrees in protein-protein interaction network of TFs in top layer with those in lower layers, and calculated the significance using the Student t-test.

Expression levels of all *C. elegans* genes at 8 different tissues at L2 stage were measured using tiling arrays. The tissues are defined as in Table S3. Tissue specificity score for a gene was calculated as follows: $TSPS = \sum_{i} f_i log_2(f_i/p_i)$, where f_i is the ratio of the gene expression level in tissue i to the gene's sum total expression level across all tissues, and p_i =1/8 for all tissues, is the fractional expression of a gene under a null model assuming uniform expression across tissues. A greater tissue specificity score suggests more specific expression in a single or multiple tissues, whereas a score of zero suggests uniform expression. Apart from tissue specificity, the stage specificity score of a gene throughout its developmental time course is defined in a similar fashion.

The *C. elegans* protein-protein interaction data were downloaded from the Worm Interactome Database (*77*). The data contain 178,152 interactions that were determined by a combination of: yeast-two-hybrid experiments, literature curation and by computational analysis.

D.4. More Detail on an "Integrated miRNA-TF Network and its Motifs"

D.4.a. Identification of Conserved miRNA Binding Sites in 3'UTRs

We made new predictions of candidate miRNA binding sites in *C. elegans* mRNAs using the integrated transcript models. Overall, we identified a total of 20,427 predicted target sites within 4,866 3′UTRs for 2,244 genes. The target sites are conserved in *C. briggsae*. (In order to identify this conservation, we use genome alignments between *C. elegans* and *C. briggsae*, and extracted the alignments corresponding to annotated 3'UTRs in *C. elegans*. Within those aligned regions, we identified target sites for *C. elegans* that are conserved in the aligned *C. briggsae* sequences (see Sect. F).

In more detail, we used the PicTar algorithm (*78*) to identify conserved microRNA target sites within annotated 3'UTRs from the aggregate integrated transcripts model (Table S13 and (*36*). We applied the version of PicTar described in (*79*) with the slight modification that a perfect seed site if covering the first 5' base of the miRNA was required to match an adenosine at this position. We used a non-redundant subset of 3'UTRs, considering only those which do not overlap any CDS in an alternative transcript isoform for the same gene, and excluding a small subset of transcripts (~4,500) for which we identified more than one putative ORF in different reading frames. We used 183 miRNAs, either annotated in miRBase14 (*59*) or newly identified from *C. elegans* embryos (*56*) using miRDeep version 2 (*44*), and genome alignments between three (*C. elegans, C. briggsae,* and *C. remanei*) or five (also including *C. brenneri* and *C. japonica*) species. This set of predictions for the aggregate integrated transcripts model are an alternative to our recently published predictions for the *C. elegans* 3'UTRome (*17*), which use 3'UTRs for AceView (*80*) gene models.

We also independently searched for perfect Watson-Crick complementary seed sites covering the first or second 5' miRNA heptamer which are prefectly conserved. These predictions should be identical to the 'TargetscanS' predictions (*81*) and, by definition, are identical to the vast majority of PicTar predictions. Indeed, a comparison of the results between the two algorithms revealed that PicTar identified 99% of seed sites predicted by TargetScan, and conversely, TargetScan identified 89% of seed sites predicted by PicTar. The reasons for the additional PicTar predictions are (i) PicTar uses a more general definition of 'conserved seed site', allowing for evolutionary changes between the different heptamers in the same alignment, (ii) PicTar also effectively locally realigns target site candidates to overcome alignment problems, and (iii) PicTar also predicts imperfect, conserved seed sites if very significantly compensated by additional basepairings between the remainder of the miRNA and the mRNA. Previous independent comparisons of miRNA target prediction algorithms using other data sets have shown that TargetScan and PicTar are top performers in the field, and generally produce the highest overlap with experimentally determined sites ((*82*); reviewed in (*83, 84*). Compared to our earlier analysis of *C. elegans* 3'UTRs (*79*), our new prediction sets ((*17*) and this study) show a higher signal-to-noise ratio compared to synthetic miRNAs of similar composition (1.8-2.4 and 2.1-3.4 for 3-way and 5-way alignments, respectively, using the method described in (*78*)). We attribute this to a combination of better multi-species genome alignments and exclusion of genomic sequence regions that are not supported by experimental evidence (previous predictions used up to 500nt downstream of the CDS when no annotated 3'UTR was available).

D.4.b. Calculation of Overrepresented Motifs in the Integrated Network

In order to identify the patterns in the integrated network that are more frequent than by chance, we enumerated all the possible patterns with 3 nodes. The frequencies of these patterns in the real network were compared with those in 1,000 random networks. The random networks were generated by rewiring the real network, while keeping its topological statistics constant; i.e., keeping the same number of coding gene targets and the number of miRNA targets for a TF node, the number of regulatory TFs and targets for a miRNA node, and the number of regulatory TFs and miRNAs for a target gene node. For each pattern, a *z*-score was calculated as follows: $Z-score=\frac{N_{real}-Mean(N_{rand})}{SD(N_{rand})}$, where N_{real} and N_{rand} are the number of corresponding patterns in the real network and in the random networks respectively. A pattern in the integrated network with a significant positive *z*-score indicates over-representation, whereas a significant negative one indicates under-representation. The *p*-value for a *z*-score was calculated by referring to a standard normal distribution. For the network motif analysis, we only used the proximal targets (500bp upstream to 300bp downstream).

E. More Detail on "Chromatin Organization and its Implications"

E.1. More Detail on "Models Relating Chromatin to TF Binding"

For each TF binding experiment, the bins that overlap with the binding peaks form the positive set. The same number of other bins was randomly sampled from the whole genome as the negative set. Half of the bins in the positive and negative sets were used as training examples to train support vector machine (SVM) models using default parameters in Weka (*85*). The other half was used to test the performance of the SVM models. Model accuracy was evaluated using ROCs, as well as the area under the ROC curves (AUROC). We also used precision-recall (PR) curves as a secondary measure, and arrived at the same general conclusions. Different feature sets were used in different configurations. Each of the single-feature models involves only one feature. The integrative model involves all features, and the stage-specific models involve only features from one development stage.

E.2. More Detail on "Models Relating Chromatin to Gene Expression"

The *C. elegans* genome was divided into bins of 100 bp. For each bin, the average signal was computed for each chromatin feature and for each TF binding experiment. Consequently, each experiment is associated with a vector of signals. Correlations were computed as the pairwise Pearson correlations between these vectors. We also computed Spearman correlations and normal-score correlations between the vectors. The correlation patterns are similar for the three correlation functions, and we include only the results based on Pearson correlations.

F. More Detail on "<u>Conservation Analysis</u>"

F.1. Multiple Alignments

The following is an extract of the information located at (*86*), which describes the methods used to build the six-way nematode alignments. This URL also provides links to downloadable files containing the nematode sequences, six-way alignments and conservation scores produced by this analysis.

The nematode sequences analyzed here were obtained from the following sources: *C. briggsae*: Washington University at St. Louis School of Medicine Genome Sequencing Center (WUSTLGSC) version 1.0, January 2007 (*87*); *C. remanei* WUSTLGSC version 15.0.1 May 2007 (*88*); *C. brenneri:* WUSTLGSC version 6.0.1 February 2008 (*89*); *P. pacificus:* WUSTLGSC version 5.0 February 2007 (*90*); *C. japonica:* WUSTLGSC version 3.02 March 2008 (*91*).

Pairwise alignments with the *C. elegans* genome were generated for each species using blastz from repeat-masked or window-masker masked genomic sequence. Pairwise alignments were then linked into chains using a dynamic programming algorithm that finds maximally scoring chains of gapless subsections of the alignments organized in a kd-tree. The scoring matrix and parameters for pairwise alignment and chaining were tuned for each species based on phylogenetic distance from the reference. High-scoring chains were then placed along the genome, with gaps filled by lower-scoring chains, to produce an alignment net.

The resulting best-in-genome pairwise alignments were progressively aligned using multiz/autoMZ to produce multiple alignments. The multiple alignments were post-processed to add annotations indicating alignment gaps, genomic breaks, and base quality of the component sequences.

F.2. Evolutionary Constraint Calculations

Conservation scoring was performed using the PhastCons package, which computes conservation based on a two-state phylogenetic hidden Markov model (HMM) (*92*). PhastCons measurements rely on a tree model containing the tree topology, branch lengths representing evolutionary distance at neutrally evolving sites, the background distribution of nucleotides, and a substitution rate matrix. Conserved and non-conserved 6-way tree models were constructed from the information in (*93*) with the branch length for *P. pacificus* arbitrarily set manually for the phastCons starting-tree model. The branch lengths in the conserved and non-conserved tree models were produced by the phastCons tuning steps using *phyloBoot*. The phastCons parameters used for the conservation measurement were: expected-length=15 and target-coverage=0.55.

F.3. Coverage Analysis

F.3.a. Overview

Because genomic elements frequently overlap, when we assigned fractions of the whole and constrained genome to element classes we must define an order in which we partitioned the genome among the elements. For the purposes of this analysis, we partitioned the genome into regions covered by gene transcription annotations including coding exons, UTRs, and ncRNAs. Regions covered by introns were excluded from the coverage analysis except for those portions that intersected with another functional element. We also included the gene-related pseudogene annotations, transcription factor binding sites, chromatin associated factors, and dosage compensation factors. The result of this choice is to assign a smaller portion of the genome to annotations added at the end than they would receive if the order were reversed. Fig. S44 gives a detailed representation of how the proportions of the whole and constrained genome are covered by elements, by showing both the unique coverage of each element class as well as the amount that overlaps with previously-added elements whereas Fig. 8A shows only the former.

F.3.b. Use of GSC Statistic

For calculating the confidence intervals for the proportion of annotated regions expected to contain constrained regions by chance, we used the Genome-Structure-Correction (GSC) statistic (*94, 95*), which corrects for internal correlations of size and position within the annotations and within the constrained regions.

F.3.c. Calculating the Genomic Coverage of TF-binding sites

The ChIP-seq technique used to identify transcription factor binding sites produces peaks that are broader than the true physical binding site. The width of the peaks depends on a number of technical factors, the chief of which is the average size of the chromatin fragments used for immunoprecipitation following experimental shearing. Complicating this is the fact that several transcription factor-binding sites may be located close to one another, resulting in broad peaks that contain several subpeaks or “summits.”

During the genomic coverage analysis, we did not wish to overestimate the coverage of the genome by transcription factor binding sites. However, we felt that the minimalist approach, which is to use the size of the TF binding recognition motif, was too extreme. The motifs are on the order of 8-12 bp in length, but the TF typically binds to chromatin in association with a complex of cofactors, and the actual chromatin-associated region may be much larger.

Our pragmatic approach was to call TF binding sites using PeakRanger, which is a refinement of the PeakSeq algorithm (*61*) that uses coverage signal topography to accurately identify summits within a broad peak (see section D.1.a). We then validated this choice by preparing profile plots of each TF experiment in which the mean PhastCons evolutionary conservation score was plotted against the distance from the center of each peak. As shown in Fig. S46 for three representative TFs, the constraint score reaches its maximum within 0-20 bp of the center of the peak, and reaches its half maximal level at roughly 100 bp from the center. On this basis, we used the peaks derived from this algorithm for all genomic coverage calculations described in the text.

G. More Detail for the "Discussion": Comparing Human and Worm Annotation

In order to compare the results obtained in *C.elegans* by the modENCODE Consortium against those obtained by the Human ENCODE Pilot Project (*94*), we compared the amount of transcription and binding by transcription factors between the *C. elegans* and human genomes. We further compared aggregation plots for both RNA Pol II and matching histone modifications. Finally, we compared the amount of conserved bases that can be experimentally annotated between *C. elegans* and human.

G.1. Analysis of the Amount of Transcription and TF Binding

For two representative samples from the ENCODE pilot (Placental and HeLa PolyA RNA) and modENCODE (L2 PolyA RNA) projects, we compared the transcribed genomic fractions (Table S16). For both ENCODE and modENCODE, transcription was detected using tiling arrays. We also examined the amount of transcription in genic regions (exonic and intronic) and intergenic regions and observed a similar percentage of intergenic transcription (15.8% (Placenta) and 44.0% (HeLa) for human ENCODE and 15.6% for *C. elegans* modENCODE), consistent with significant amounts of novel intergenic transcription in both species. We used GENCODE annotation for human and WormBase WS190 annotation for *C. elegans.* Similarly, for a number of representative transcription factors we compared the amount of genic versus intergenic binding between ENCODE and modENCODE. We selected the following ChIP-chip datasets from the ENCODE Pilot Project: STAT1, cFos, cJun, CTCF and CEBPe (the first three were performed in HeLa cells and the last two in HL60 cells). For the modENCODE project we selected CEH-14 (L2), EGL-27 (L1), MAB-5 (L3), PES-1 (L4) and PHA-4 (EMB). We first observed that for both human and *C. elegans*, the majority of transcription factor binding occurs in intergenic regions (88% for human and 91% for *C. elegans*) (Table S15). We also observed that the fraction of intergenic sequence that is bound for each transcription fraction is lower for *C. elegans* compared to human (1.11% versus 1.80%). Given the significant difference in the sizes of the genomes, it is not clear *a priori* whether or not we expect to see a greater fraction of intergenic binding in human. A complicating factor is that the human TF binding experiments were performed using ChIP-chip, which has lower resolution than ChIP-seq.

G.2. Aggregation Analyses: RNA Pol II and Histone Modifications

We first compared aggregation plots of RNA Pol II ChIP-Seq signal around human TSSs and *C. elegans* TSSs. For human, we used published RNA Pol II ChIP-Seq data from HeLa cells (*61*) which was assayed as part of the whole-genome phase of the ENCODE Project. We compared this dataset against *C. elegans* RNA Pol II ChIP-Seq performed in the L4 stage of development. CCDS (*96*) TSSs were used for human and TSSs from WormBase WS180 (*97*) were used for *C. elegans*. In Fig. S49 we see that the normalized aggregation plots look very similar. Similarly, aggregation plots for histone modifications common to both modENCODE and ENCODE Pilot phase were generated over both TSS and TTS. Fig. S50 (drawn to be comparable to Fig. 6) shows data from (*98*) based on NGS ChIP-Seq data. The signal values are from the NPS algorithm used to process the ChIP-seq data at nucleosome resolution (*99*). Unlike the case of Pol II, the histone marks appear different in *C. elegans* and human.

G.3. Conservation Analysis

In comparison to the evolutionary constraint analysis published by the Human ENCODE Pilot Project (*94*), the region of the *C. elegans* genome under purifying selection is much larger (29.6% vs 4.9%, Fig. S48). This finding reflects both the compact nature and the higher proportion of coding vs. noncoding regions in the nematode genome. There are also differences in how the various classes of functional elements contribute to the constrained portion of the genome. The biggest difference is the amount of unannotated constrained bases, which was 40% in the ENCODE pilot, and about half this value (20.5%) in modENCODE (Fig. S48). This difference is almost entirely due to the proportion of constrained bases covered by coding exons. In the ENCODE pilot, 32% of the constrained portion of the genome was attributable to coding exons, while in modENCODE, over 53% of constrained bases are coding. Other annotations, including UTRs and annotations of classes involved in transcriptional regulation and chromatin maintenance, are in similar proportions in ENCODE versus modENCODE. Hence, we can infer that C. *elegans* constrained genome contains a substantially higher proportion of coding bases than the regions sampled by the ENCODE pilot, and speculate that C. *elegans* may have a smaller proportion of its genome devoted to regulation, chromatin maintenance, or other functions. However, our ability to extrapolate from the ENCODE pilot to the whole human genome is limited by the small amount of the human genome that was sampled in the pilot (1%) and the fact that the ENCODE pilot's choice of human genomic regions to be analyzed was not entirely random. In addition, a recent study's re-estimate of the proportion of the human genome under evolutionary constraint was revised upwards to 6.5-10% (*100*), which will also affect the interpretation of the ENCODE results. A full comparison will have to await the publication of the full ENCODE data set.

H. Author Roles

Data generation:

Julie Ahringer, Cathleen M. Brdlik, Jennifer Brennan, Jeremy Jean Brouillet, Ming-Sin Cheung, Luke O. Dannenberg, Abby F. Dernburg, Arshad Desai, Lindsay Dick, Andréa C. Dosé, Thea Egelhofer, Sevinc Ercan, Ghia Euskirchen, Brent Ewing, Reto Gassman, Ting Han, Steven Henikoff, LaDeana W. Hillier, Heather Holster, Tony Hyman, David M. Miller III, Kohta Ikegami, A. Leo Iniguez, Judith Janette, Morten Jensen, Masaomi Kato, Vishal Khivansara, John K. Kim, Stuart K. Kim, Paulina Kolasinska-Zwierz, Isabel Latorre, Amber Leahey, Jason D. Lieb, Michael MacCoss, Marco Mangone, Gennifer Merrihew, Andrew Muroyama, John I. Murray, Wei Niu, Hoang Pham, Taryn Phippen, Fabio Piano, Elicia A. Preston, Valerie Reinke, Heidi Rosenbaum, Mihail Sarov, Frank J. Slack, Cindie Slightam, Michael Snyder, William C. Spencer, Susan Strome, Teruaki Takasaki, Dionne Vafeados, Anne Vielle, Ksenia Voronina, Guilin Wang, Robert H. Waterston, Christina Whittle, Beijing Wu, Mei Zhong, Xingliang Zhou

Data analysis:

Ashish Agarwal, Roger P. Alexander, Pedro Alves, Bradley I. Arshinoff, Raymond K. Auerbach, Galt Barber, Adrian Carr, Aurelien Chateigner, Chao Cheng, Hiram Clawson, Sergio Contrino, Jiang Du, Xin Feng, Mark B. Gerstein, Phil Green, Francois Gullier, Kristin C. Gunsalus, Michelle Gutwein, Lukas Habegger, Jorja G. Henikoff, Stefan R. Henz, LaDeana W. Hillier, Angie Hinrichs, W. James Kent, Ellen Kephart, Ekta Khurana, Stuart K. Kim, Jing Leng, Suzanna Lewis, Tao Liu, X. Shirley Liu, Paul Lloyd, Lucas Lochovsky, Yaniv Lubling, Zhi John Lu, Rachel Lyne, Sebastian D. Mackowiak, Sheldon McKay, Desirea Mecenas, Gos Micklem, Mitzi Morris, Eric L. Van Nostrand, Siew-Loon Ooi, Marc Perry, Nikolaus Rajewsky, Gunnar Rätsch, Andreas Rechtsteiner, Kahn Rhrissorrakrai, Rebecca Robilotto, Joel Rozowsky, Kim Rutherford, Peter Ruzanov, Rajkumar Sasidharan, Andrea Sboner, Paul Scheid, Eran Segal, Hyunjin Shin, Chong Shou, Richard Smith, William Clay Spencer, Lincoln Stein, E.O. Stinson, Scott Taing, Nicole L. Washington, Koon-Kiu Yan, Kevin Y. Yip, Georg Zeller, Zheng Zha

Lab project coordinators (PIs and co-PIs):
Julie Ahringer, Abby F. Dernburg, Arshad Desai, Mark B. Gerstein, Phil Green, Kristin C. Gunsalus, Steven Henikoff, LaDeana W. Hillier, Tony Hyman, David M. Miller III, A. Leo Iniguez, W. James Kent, John K. Kim, Stuart K. Kim, Eric C. Lai, Suzanna Lewis, Jason D. Lieb, X. Shirley Liu, Michael MacCoss, Gos Micklem, Fabio Piano, Nikolaus Rajewsky, Valerie Reinke, Eran Segal, Frank J. Slack, Michael Snyder, Lincoln Stein, Susan Strome, Robert H. Waterston

Project management:
Elise A. Feingold, Peter J. Good, Mark S. Guyer, Rebecca F. Lowdon

I. Acknowledgements

Funding from the National Human Genome Research Institute of the NIH as part of the modENCODE project, NIH (grant R01GM088565), Muscular Dystrophy Association and the Pew Charitable Trusts (JKK); the Helmholtz-Alliance on Systems Biology (Max Delbrück Centrum Systems Biology Network) (SDM); the Wellcome Trust (JA); and the AL Williams Professorship funds (MG). We appreciate help from S. Anthony, K. Bell, C. Davis, Y. Field, A. S. Hammonds, J. Jo, N. Kaplan, A. Manrai, B. Mathey-Prevot, R. McWhirter, S. Mohr, S. Von Stetina, J. Watson, K. Watkins, C. Xue, and Y. Zhang, and B. Carpenter, C. Jan and D. Bartel for sharing data on polyA sites prior to publication, WormBase curator G. Williams for assistance in quality checking and preparing the transcriptomics datasets for publication, as well as his fellow curator P. Davis for reviewing and hand checking the list of pseudogenes.

Supplementary Figures

Fig. S1: ChIP-chip and ChIP-seq comparison

A. Pairwise Scatter plot and correlation between Pol II ChIP-chip and ChIP-seq replicates with combined profiles at two developmental stages (early embryo, left and L4, right). The sample names are shown on the diagonal. In the lower triangular part of the panel, each blue dot represents the median signal levels of ChIP-chip (MA2C score) and ChIP-seq (sequence read count) within a 1kb-segment on the genome. The upper triangular part provides the correlation coefficient of each pair.

B. The heatmap image represents pairwise correlations between ChIP-chip and ChIP-seq combined profiles at early embryo and L4 stages, and is hierarchically clustered by both rows and columns. It is shown that the variation between the two platforms at the same stage (correlation coefficient of about 0.7) is smaller than that between the two different stages of the same platform (correlation coefficient of 0.4-0.53).

C. The venn diagrams show the overlap of the top 3000 Pol II binding sites identified by ChIP-chip (blue circle) and ChIP-seq (red circle) in early embryo (left) and L4 (right) stages. It can be seen that more than 2/3 of Pol II binding sites were commonly identified by the two platforms.

Fig. S2: Correlation of RNA expression levels for young adult between RNA-seq and tiling array platforms

Each data point represents a gene. To account for multiple isoforms, a gene is here defined as the union of all exonic nucleotides. RNA-seq expression levels per gene were measured using RPKM, and tiling array levels were measured using the mean intensity of probes falling within exons. The genes in the upper left likely represent cross-hybridization in tiling arrays.

Fig. S3: Numbers of RNA-seq reads

Total reads along with numbers of uniquely and multiply aligned non-rDNA reads for each of the 19 *C. elegans* stages and samples. Total reads are defined as those that passed the Illumina quality filters. The largest proportion of non-uniquely aligned reads are those aligning to rDNA regions of the genome.

Fig. S4: RNA sequencing depth analysis

A. Density plots of the expression of 20,051 genes in WormBase190. Each line corresponds to a sequencing depth. The legend reports the number of mapped reads (in millions). The two peaks represent genes not expressed (left) and expressed (right) at each sequencing depth. Note that the number of non-expressed genes drops sharply at first as sequencing depth increases, then reaches a plateau.

B. Pair-wise comparison of the density plots. Y-axis reports *p*-values of the Kolmogorov-Smirnov test as a function of depth of sequencing (x-axis). The dotted line shows a *p*-value of 0.01. Higher *p*-values (>0.01) indicate no significant difference between the distributions. The plot shows that a sequencing depth between 13.4 and 16.8 million reads is sufficient to capture most expressed genes in whole animal samples.

C. Rate of gene discovery. The number of genes with RPKM=0 are reported as a function of sequencing coverage. The equation reports the coefficients and the $R^2$of the best fitting exponential curve. The fitted curve is: Number of non-expressed genes = 8.5 x (depth of sequencing) $^{-0.88}$(R^2=0.90).

Fig. S5: Transcript building

This diagram illustrates the process of gene model construction. The top half shows the various features identified through RNA-seq and the bottom half shows the resultant models. To build gene models in regions across the genome we search for the most abundantly represented splice junction, indicated by "(1)", and then move away in both directions until another feature is encountered. Moving to the right in this example, coverage continues until a second splice junction is encountered, so the model incorporates this junction and continues through the next area of coverage until the end of coverage is encountered. Here, this position corresponds to a polyA site, indicating a transcript stop signal (black line). Moving to the left of the initiating splice junction, a splice junction is again encountered and incorporated. The first gene model is completed when the end of coverage is encountered. A splice junction indicated by "(2)" that was not incorporated into the first model is then used to initiate a second gene model. Moving to the right, this gene model is the same as model 1. Moving to the left, it encounters the end of coverage, with an associated start site (either a spliced leader junction or a strand bias signal) and the model is complete. Orientation is implicit in the sequences of the splice junctions and the start and stop sites.

Fig. S6: A complex isoform example

This region of the transcript ZK783.1, homologous to human fibrillin-1, illustrates that alternative splicing in *C. elegans* can be quite complex. The current WormBase model (WS190) is shown at the top with our aggregate integrated transcript models shown below. Raw read counts per base for early embryo (orange) reveal clearly evident splice junctions, whereas in L3 (blue gray), a series of introns are apparently read through without splicing until splicing to either the penultimate exon in the region or skipping this to the final exon shown.

Fig. S7: Features defined by RNAseq as compared to WormBase as of January, 2007 (WS170)

Number of features identified by stage as compared to features in WormBase (WS170) when the modENCODE project began. The two right most bars represent the RNA-seq-only aggregate set and the aggregate integrated transcript set created from all available *C. elegans* transcriptome data. All features (TSS is Transcript Start Site, SL1 and SL2 are Spliced Leader sites) were clustered when within 25 bases of one another. For example, if there were three different polyA sites within 25 bases of one another, they were counted as a single polyA site.

Fig. S8: Number of confirmed splice junctions over time

This figure indicates the significant contribution of RNA-seq to annotating the *C. elegans* genome. There were 11,467 splice junctions confirmed when the complete *C. elegans* genome sequence was first published (*101*). The first rise in 2003 was a result of the OST Project (*102*) and the remaining increases were a result of the modENCODE project (e.g. (*14*)). The number of RNA-seq datasets added at each time point is indicated.

Fig. S9: Proportion of splice junctions confirmed by various methods

The large overlap in splice junctions confirmed between RNA-seq, RT-PCR/RACE and mass spectrometry (*16*) provide confidence in the methods used for identifying confirmed junctions by RNA-seq.

Fig. S10: Saturation of discovery of additional ncRNAs and coding exons with additional RNA-seq data sets

We are presently utilizing a number of approaches to ncRNA discovery, and our initial efforts have revealed thousands of new ncRNAs from the *C. elegans* genome. As assays are performed under additional conditions, as we refine our computational methods of analysis, we expect to discover many thousands more ncRNAs. The saturation plot for novel ncRNAs (left) illustrates this point. In each experimental condition, the total length of ncRNAs expressed was determined using a combination of experimental and computational methods. When multiple conditions are considered together, the total length of ncRNAs depends on the set of conditions involved. The saturation plot displays that total length (y-axis) at different number of conditions (x-axis). At each point along the x-axis, all possible combinations of conditions are considered, and the distribution of total lengths is summarized by a box plot. The black line shows the slope of the curve connecting the averages at the end of the curve. The steepness of the curve suggests that more ncRNAs are expected to be discovered if additional conditions are considered. We made the same saturation plot (on the right) for coding exonic regions. The detection of expressed exons tend to be saturated when additional experiments are added.

Fig. S11: Number of stages and samples where a given gene or splice junction is observed

Most genes and splice junctions are represented in all 19 stages and conditions, with smaller peaks for those found in only one or two stages and samples. The peak at 2 for stages per gene/splice junction in part results from the requirement that all novel splice junctions occur in at least two different stages (novel is defined as not a part of WormBase170 predictions, which included WormBase, Twinscan and Genefinder predictions).

Fig. S12: Developmental stage-specific expression

A. Expression profiles of developmental stage-specific genes. High and low expression levels (normalized DCPMs) are shown in red and blue, respectively. Expression levels of each gene are normalized across the 7 developmental stages by subtracting the mean then dividing the standard deviation.

B. Expression profiles of the meta-genes for developmental stage-specific transcripts**.** The expression level for a meta-gene was calculated by averaging the expression levels of all genes which are specific to a given developmental stage.

C. Enrichment of promoter motifs. Enrichment of 24 EE-specific candidate motifs identified by the MEME algorithm in promoters of stage-specific genes. The –log(*p*-value) was calculated by comparing the occurrences of a motif in stage-specific transcripts relative to all the other transcripts, and then color-coded with red (indicating over-representation) or blue (indicating under-representation).

Fig. S13: Lab batch effects

Distribution of Values Along Principal Component 1 By Lab. No significant batch effect due to lab was found (p=.068, t-test)

Fig. S14: Cumulative plot of isoform composition distribution

This histogram shows the distribution of differences in isoform composition for all genes with multiple isoforms (12,875) in 21 pairwise comparisons across 7 developmental stages (EE, LE, L1, L2, L3, L4, YA). Isoform composition for gene *i* in stage *S* is represented by a vector $\theta(i,S,k)$ where the kth component is the relative abundance of isoform *k* in relation to the other isoforms. Between two stages *R* and *S* for a given gene i the difference in abundance vectors gives a measure of the change in isoform usage for a gene. This is represented as $D(i,R,S) = \sum/k \ ((\theta(i,R,k) - \theta(i,S,k))2)/k$. The difference *D* is a fractional number between 0 and 1; scores close to 1 indicate dramatic differences in the relative composition of different isoforms of the gene. The histogram plots the distribution of *D* values for all genes *i.* It is averaged over all pairs of stages *R* and *S*. The error bars represent the range of number of genes in every histogram across the 21 pairwise comparisons. Overall, the histogram shows that most genes have minor differences in their isoforms, but a small fraction (280) have major and minor isoform switching between stages. (The minor isoform is defined as that with the lowest expression and account for less than 15% of the total expression of a given gene, while the major isoform account for more than 85% of total expression). This is a cumulative version of Fig. 1B.

Fig. S15: SOM clusters of transcripts with different developmental expression profiles

Application of a Self Organizing Map (SOM) algorithm to developmental transcript expression profiles for 15 stages resulted in 48 different SOM clusters across development. Individual transcript-level expression was calculated based on probabilistic inference using the deepseq9 algorithm. The $\log_2$ value of probabilistic read counts from deepseq9 (y-axis) is plotted for each of 15 developmental stages (x-axis) arranged in the following order: MxE (male *him-8*), EE, LE, L1, L1 (lin-35), L2, L3, dauer entry, dauer, dauer exit, L4, L4 male, L4 soma, YA, and aged adult (*spe-9*). All dauer stages are *daf-2*. Solid lines represent mean transcript expression in each of the 15 staged samples; dashed lines represent one standard deviation from the mean. (Data shown are for 15 discrete timepoints and do not represent a continuous change in expression across intermediate timepoints.)

Fig. S16: Number of genes and transcripts shared between pairs of SOM clusters

Adjacent to each cluster ID (c1..c48) is its size, indicated in terms of genes (yellow, g=XX) or transcripts (green, t=XX). Cells are shaded by the Pearson Correlation Coefficient (PCC) between developmental expression profiles for each pair of clusters, calculated from their mean expression across the 15 staged samples. Values within cells indicate number of shared genes and transcripts in the yellow and green bounded regions, respectively.

Fig. S17: Classes of distinguishing features between isoforms with different developmental expression profiles based on SOM clustering

Shown are the numbers of alternative transcript pairs for the same gene that fall into different SOM clusters, for cases in which a single isoform falls into one SOM cluster and one or more alternative isoforms fall into another cluster (see text for details).

Fig. S18: Examples of read count distributions supporting differential expression of alternative transcript isoforms among developmental stages

Aggregate integrated transcript modelsfor genes with transcripts falling into different SOM clusters are displayed with wiggle plots from relevant stages using the Integrative Genomics Viewer (*103*). These plots represent 36-mer reads aligned without mismatch (trimmed up to 2 bases) and were calculated by the SHRiMP aligner v1.3 (*104*).

Forward and reverse primers used for qRT-PCR validation of transcripts shown in A and B are indicated with red and green arrows, respectively. The reverse primer for F26B1.2.T8 contains a 3' poly-dT anchor.

A. Unique 5' UTRs of T3 and T4 isoforms of C23H3.7. The T3 isoform is absent in young adult and is co-expressed with the T4 isoform in early embryo, but is not detected in young adult..

B. An alternative CDS exon is skipped in F26B1.2.T8 and included in F26B1.2.T9. The T9 isoform is more highly expressed in L4 than L2.

C. Overlapping 3' UTR of F28C6.3.T2 and F28C6.3.T4. The T4 isoform is expressed at a much higher level in L4 than in young adult.

Fig. S19: Breakdown on how the updated list of *C. elegans* pseudogenes was created

The figure schematizes the workflow in updating the pseudogenes in WormBase, to arrive at current total of 1,293 pseudogenes. Pseudogenes came from two sources: those already in WormBase annotations (right) and those identified by Pseudopipe (left). The initial overlap of 1,025 pseudogenes from these two sources was kept. The remaining subsets also kept are shown in red. These include 83 additional duplicated (DUP) and 90 additional processed (PSSD) pseudogenes identified by pseudopipe. They also include 95 pre-existing WormBase annotation not found by pseudopipe that were double-checked by the WormBase curators.

Fig. S20: Binning of long TARs built from tiling arrays

The TARs from tiling array data were built from the union of 41 samples (*11*). The minimum length of TAR is 100nt. Since long TARs could cover more than one type of sequence element, such as exons, introns, and UTRs, they were spliced into small windows of at most 150nt each, with adjacent windows having a 75nt overlap. Each bin was defined as intronic TAR, exonic TAR or UTR depending on which annotation it overlaps (WormBase170 was used). Those small TARs that are less than 150nt are not spliced.

Fig. S21: Predicting ncRNAs

A. The two panels illustrate the increased power achieved by combining features to discriminate between ncRNAs and other regions of the genome. These graphs show the distribution of expression feature values (e.g. from small RNA-seq) for genomic regions in the worm genome corresponding to ncRNAs and other types of sequence elements. The two panels show that while each feature alone cannot discriminate among different types of genomic elements, combining features into an integrated model can enable differentiation. The left panel shows the distributions of expression values for four representative features of the nine features examined using the gold-standard set of annotated regions (see (*60*) for the definition of the gold-standard set). The gold standard consists of four types of genomic elements: the known non-coding RNA, coding sequences (CDSs), untranslated regions (UTRs), and intergenic regions. A scatter plot of individual regions with values normalized to the same scale shows that the known ncRNAs are not readily distinguished from other regions, particularly using the bottom two features. At right, the maximum signal of polyA RNA on a tiling array is plotted in a two-dimensional scatter plot against predicted secondary structure conservation. Even using just two features, the ncRNAs begin to separate from the other regions. Expression values in the right panel are log-transformed normalized read counts (DCPM). Where multiple experimental data sets exist, the maximum value is used. The data used in the plots are from gold standard bins defined in (*60*).

B. Example of a novel ncRNA with support from multiple sources of information in embryos. Track labels are PHA-4, HLH-1, RNA Pol II: ChIP-seq reads from the indicated protein, where signal heights are normalized by their total mapped reads; H3K27ac (histone 3 lysine 27 acetylation), H3K4me (H3K4 methylation): log-transformed values of the ChIP-chip data for two chromatin features normally associated with active genes; PolyA and Small RNA-seq: reads from polyA-selected and small RNA sequencing; Total RNA tiling arrays: log-transformed values of transcription on the tiling array in embryo; TARs: Transcriptionally Active Regions called from the tiling array signal track; Refseq: annotated genes in the region. The grey box at center shows a novel non-coding RNA ~160 nt in length captured only by the tiling array, indicating that it is not polyadenylated and is longer than the 30 nt size cutoff of the small RNA-seq experiment.

Fig. S22: TF binding around non-coding RNAs

A. Enrichment of binding targets and signal of TFs in non-coding vs. coding genes. Max signal value represents the ratio of maximum binding signal of a TF around its target non-coding genes to that of its target coding genes. Target fraction represents the ratio of target percentage in non-coding genes to that in coding genes. Only TFs present in the larval stage samples are shown. Some factors such as GEI-11 clearly bind more to ncRNA than others (e.g. PHA-4).

B. An example showing GEI-11 binding near three ncRNAs. Four other factors (MAB-5, LIN-39, EGL-27, and PES-1) are also shown as controls. The signal for each TF, as well as for Pol II and input, plots the ChIP-seq raw read counts scaled based on total mapped reads. Pol II and input samples were from N2 animals; TF samples were from animals expressing the factor tagged with GFP. The value of tiling array ChIP-chip signal for H3K27ac and H3K4me are also shown in green. Raw reads of polyA-plus RNA-seq and small RNA-seq, as well as expression (log2 of signal) from total RNA tiling array signal are also shown. The ncRNA annotations and protein annotations are from Refseq (*105*).

C. Average ChIP-seq signal around the transcript start site (TSS) of target coding (red) and non-coding (blue) transcripts for four representative TFs. The signal is the normalized mapped reads over input at each position (window size is 100nt).

Fig. S23: Co-occurrence of transcription factors

Co-occurrence is counted if two TFs bind to the promoter region of the same gene (2000 bp upstream to 300bp downstream of TSS), without considering the strength of binding. Genes targeted by HOT regions were removed before calculating the co-occurrence. The heat map reflects the co-bound correlation of each pair of TFs at targeted gene loci, with red indicating more co-bound genes than would be expected by chance and blue, indicating less. TFs have been clustered along both axes based on the similarity of their bound targets with other factors. The same stage is annotated with the same color. The HOX genes are highlighted with orange color.

A. Co-occupancy of transcription factor pairs at targeted coding genes.

B. Co-occupancy of transcription factor pairs at targeted miRNAs.

Fig. S24: Comparison of PeakSeq and PeakRanger peak calls

The PeakRanger algorithm refines the PeakSeq peak calling package by identifying narrow, highly specific summit peaks within the broad regions called by PeakSeq. This figure illustrates PeakRanger's performance across a representative 12 kb region of chromosome I from a PHA-4 early embryo ChIP-seq experiment. From the top, the tracks are: (1) ChIP-seq signal graph; (2) PeakRanger-identified summits; (3) PeakRanger summits extended 100 bp in each direction; (4) Broad regions captured by unmodified PeakSeq algorithm.

Fig. S25: Distribution of TF binding

A. Many regions show overlap of ChIP-seq binding sites for 23 transcription factors. Red indicates the number of regions bound by 1 to 23 TFs in ChIP-seq data. There are 16,707 genomic regions bound by at least 1 TF, and 304 regions bound by at least 15 factors. Black indicates the average number of regions bound in 1,000 randomized controls, with error bars indicating standard deviation. In randomized controls, an average of less than 1 region was bound by 12 or more factors, and no regions bound by 15 or more factors were observed.

B. HOT region definitions are largely insensitive to peak width. The 304 HOT regions were initially defined using broad PeakSeq peak calls (~400 nt wide). To define narrower peaks, peak summits were identified using the PeakRanger algorithm (described in SOM), and narrower peaks were defined as the region 200, 150, 100, 50, and 26nt wide centered around a peak summit. Using identical methods as previously used to identify regions of overlap between factors, we determined the percent of HOT regions that remained bound by 15 or more TFs using the narrower peak regions, as indicated by the bar heights. More than 80% of HOT regions remained HOT when 100-nt wide peaks were used.

Fig. S26: Control experiments for HOT regions

A. The x-axis plots the percentage of the 304 HOT regions which are significantly enriched in the various ChIP-seq controls. In order to verify that the antibodies used do not bind non-specifically to GFP-tagged proteins, IgG negative control experiments were performed in two different transgenic *C. elegans* lines expressing LIN-15B::GFP or EGL-27::GFP. In order to verify that GFP-specific antibody does not pull down any other proteins in *C. elegans*, GFP antibody negative controls were performed in wild-type animals at embryonic and L3 stages. As a positive control, LIN-15B antibody was used in wild-type N2 animals to immunoprecipitate endogenous LIN-15B.

B. DPY-27 only binds to HOT regions on the X chromosome. The y-axis shows the number of HOT regions found on each chromosome. The set of all 304 HOT regions and the 298 HOT regions that are bound by LIN-15B are evenly distributed across all 6 chromosomes (with chromosomes separated by color). In contrast, all 29 HOT regions bound by DPY-27 are on the X chromosome. DPY-27 regulates gene expression specifically on the X chromosome for dosage compensation (*68*)

Fig. S27: HOT regions enrichments

A. (Left) HOT regions containing HLH-1 binding show a relative lack of HLH-1 binding motifs. In black, the frequency of the *in vitro* HLH-1 binding motif (hexamer CAGCTG) is greater in HLH-1 factor-specific regions than in HLH-1 binding sites within HOT regions. The sequences in HLH-1 peak regions were randomized using the Fisher-Yates shuffling algorithm, and motif density was calculated for these shuffled regions (grey bar, error bars indicate standard deviation). 598 HLH-1 factor-specific targets are defined as regions with 1-4 factors (including HLH-1); 165 HOT regions are bound by 15 or more factors (requiring inclusion of HLH-1). (Right) HLH-1 binding does not correlate with muscle expression in HOT regions. Genes associated with factor-specific peaks for HLH-1, a muscle-specific TF, are over 7-fold more likely to be muscle-specific genes (*69, 106*) than genes located near HLH-1-containing HOT regions. For each dataset, the frequency of muscle-specific genes is shown in black, and the frequency in random gene sets of equal size is shown in grey (error bars indicate standard deviation).

B. Factor-specific and HOT regions are different in motif frequency and tissue-specificity. Bars show enrichment for motif density or tissue-specific genes between factor-specific and HOT targets for the TF. * indicates P<0.01 and ** indicates P<0.0001 by Fisher's exact test. For four factors (HLH-1, ELT-3, MDL-1, and PHA-4) with published sequence binding motifs, factor-specific regions are significantly enriched for the frequency of motifs as compared to HOT regions. Factor-specific targets compared to HOT targets are enriched for genes with tissue-specific expression patterns. L1 muscle-specific and L4 intestine-specific genes were obtained from previous studies; embryonic tissue-specific genes for body wall muscle (b.w.m.), intestine, and hypodermis were identified from embryonic tissue-specific tiling arrays. Shown are all TFs for which factor-specific peaks were significantly enriched (> 2.5 fold-enrichment and P<10e-5) for a set of tissue-specific genes when compared to all WormBase genes.

C. Genes near HOT regions are enriched for essential function. Genes were separated based upon the presence of ChIP-seq peaks within 1kb of the TSS. The y-axis shows the percent of genes bound only by 1-4 factors ("specific targets") or genes bound by 15 or more factors ("HOT regions") that serve essential functions, as indicated by RNAi knockdown. The dotted line signifies the percentage of all genes that are essential. By Chi-square test, genes nearby HOT regions are significantly more likely to be essential (9-fold; P < 10e-40), whereas genes that only had specific peaks were not.

Fig. S28: Higher gene expression level in HOT regions

A. Higher gene expression level in HOT regions detected by RNA-seq. Genes adjacent to HOT regions have significantly (P < 10e-30 by Kolmogorov-Smirnov test) higher expression in late embryonic (LE) animals than do genes located near just 1-4 bound factors. In red, expression level of genes with a HOT region within 1kb upstream or 500nt downstream of the TSS; in blue, expression level of genes proximal only to regions bound by 1-4 factors. The histogram plots the frequency (y-axis) of genes with the listed RNA-seq expression levels in late embryonic animals (x-axis, measured as log10(depth of coverage per million reads)).

B. Higher gene expression level in HOT regions in tiling arrays. Genes adjacent to HOT regions have higher expression levels in tissue-enriched tiling arrays across all tissues assayed. In red, expression level of genes with a HOT region within 1kb upstream or 500nt downstream of the transcription start site; in blue, expression level of genes proximal only to regions bound by 1-4 factors. The histogram plots the median of normalized gene expression measurements (y-axis) of genes on the listed tiling array experiment (*x*-axis), with error bars indicating standard error of the mean. Data is further described in (*21*). For embryonic experiments (left), tissue-specific gene expression measurements were obtained from tiling arrays performed on FACS sorted cells expressing a tissue-specific GFP label. For post-embryonic experiments (right), gene expression measurements were obtained from tiling arrays performed on samples that were tissue-enriched using the mRNA tagging method. In all experiments shown, genes adjacent to HOT regions are significantly shifted towards higher expression (P < 10e-15 by Kolmogorov-Smirnov test).

C. HOT genes are highly and ubiquitously expressed across tissues and developmental stages. HOT genes are identified as those with at least one HOT regions in their promoter region (from1kb upstream to 500bp downstream of TSS). HOT regions are defined as those bound by at least k TFs (k=1, 2, ...20). k=0 corresponds to the whole set of genes. Stage specificity score and tissue specificity score are calculated as described in SOM D.3.

Fig. S29: HOT regions are broadly expressed

Single-cell gene expression measurement of promoter transcriptional reporter constructs in L1 animals from 3D confocal data stacks (data from (*107*)). The x-axis represents 363 specific cells present in the L1 stage, and the y-axis shows expression of 93 mCherry reporters, with the expression level of the mCherry reporter shown by the red scale bar. Promoters containing HOT regions (bound by 15 or more factors), and even promoters containing regions bound by 10-14 factors, show broad expression across 363 cells in the L1 stage, whereas promoters lacking these regions show a variety of diverse tissue-specific expression patterns. Data is presented identically to Fig. 3C, and gene names are provided in addition to row label codes from Fig. 3C.

Fig. S30: Pair-wise correlations of PHA-4 binding signal across different stages

The union of all PHA-4 binding sites were merged together and then binned into 100nt windows. The raw reads of ChIP-seq data for each window were calculated and normalized over the respective input for each ChIP-seq experiment. The correlation coefficient of each pair of stages was then calculated. HOT regions were removed before merging the binding sites.

Fig. S31: Examples of Pol II binding and expression

Heatmap showing percentage of RNA Pol II binding and expression for *isl-1* and *pgp-2* and C15F1.2, during seven stages of the *C. elegans* life cycle. For each transcript, RNA Pol II binding levels and gene expression levels increase in concert until the stage where both reach maximum levels. In the following stages, the expression levels tend to drop at a faster rate than RNA Pol II binding. These examples illustrate one scenario in which a change in gene expression in earlier stages can be predictive of a similar change in RNA Pol II binding levels during later stages. The heatmap is normalized independently along the columns, with the values representing the ratio of signal in a stage to the maximum signal observed.

Fig. S32: Histone marks distribution over repetitive elements

Five repetitive element classes were extracted from WormBase190. The region of the genome underneath each element was subdivided into 10 equal sized bins centered on the element. In addition, the 1 kb regions flanking each element were subdivided into an additional 20 100 bp bins. The mean *z*-scores for ChIP-chip chromatin marks from L3 larvae were then graphed across each bin. The histone marks from top to bottom are: H3K27ac, H3K36me2, H3K36me3, H3K4me2, H3K4me3, H3K27me3, H3K9me1, H3K9me2 and H3K9me3.

Fig. S33: Promoters of chromosome X genes have higher GC content compared to autosomes

A. Average GC content is plotted for chromosome X and autosomal genes centered at their transcription start sites(GC content is calculated within 25 bp upstream and downstream of each coordinate). A region between -250 to -50 shows a spike in GC content on chromsome X.

B. Distribution of average GC content within this region is plotted. Chromosome X gene promoters have significantly higher GC content, as determined by a Wilcoxon rank sum test (P <2.2e-16).

Fig. S34: Histone marks aggregation around TSSs and TTSs

Average gene profiles around the TSS and TTS of various histone marks displayed for the X chromosome (red), and autosomes (blue). Genes were further stratified according to their expression level, with the top 20% of expressed genes shown in darker shade, and the bottom 20% of expressed genes shown in lighter color. The top two panels show that histone variant H3.3 marks regions of active chromatin on both autosomes and the X chromosome. _Marks typically associated with active or repressed transcription are labeled on the left. Plots from L3 stage animals (bottom row), highlight some of the differences in histone mark patterns between the EE and L3 stages. For example, H3K27me1 and H3K27me3 show stronger enrichment of expressed genes on the X in EE, whereas H4K20me1 is more strongly enriched on the X in L3.

Fig. S35: TF sequence motif discovery

A. Recovered motifs. Transcription factor ChIP-seq peak data sets were searched for enriched motifs as described in the text . Of the 23 data sets analyzed, enriched motifs were found in 22; however, only 8 transcription factors showed sufficient specificity to be accepted (see panels B and C for example). We additionally found three motifs significantly enriched in HOT regions, the first of which is partially identical to a SLR-2-reponsive motif found in wormbase.

1- Also enriched in HOT regions. The fact that the CEH-14 motif is enriched in HOT regions either means that this TF binds specifically to HOT regions, or that this TF has a weak motif and that the observed motif is derived from another protein co-binding in HOT regions. Additional experiments will be necessary to decide between these two cases.

2- Consistent with a previously published motif for the given TF.

B. Example motif distribution analysis (BLMP-1)- Distribution of BLMP-1 motif "TTTCACTTT" was plotted relative to SPP-point-binding positions (single-base-pair genomic coordinates with highest likelihood for binding (*63*)) for BLMP-1. The motif occurrence distribution is Gaussian-like around BLMP-1 point binding positions (black and yellow lines) while relatively evenly distributed over random upstream regions (red line). Black indicates high confidence peaks with SPP assigned FDR <0.01. Yellow indicates low confidence peaks with SPP assigned FDR >0.01 and <0.05. Red indicates random upstream regions.

C. Example motif density analysis for BLMP-1. 200 base pairs flanking point binding positions for BLMP-1 were analyzed for density of BLMP-1 motif "TTTCACTTT" in occurrences per base pair. BLMP-1 peaks have significantly higher occurrences of the motif than any other transcription factor. Random upstream regions and HOT regions were also analyzed on a motif-per-nucleotide scale and similarly show much lower motif density than what is found in BLMP-1 peaks.

Fig. S36: TFs in the larval network

Names of the TFs in the network in Fig. 4C.

Fig. S37: Network motifs

Three over-represented motifs in the integrated miRNA-TF network in Fig. 4A: TF (triangle), miRNA (circle), target gene (square). P-values are calculated based on an ensemble of rewired networks.

Fig. S38: TF binding and chromatin features

A. Correlations between whole-genome transcription factor binding signals and chromatin features. The Pearson correlation between the signals from each of the 27 transcription factor ChIP-seq experiments (rows) and 22 chromatin features (columns) across the whole genome are shown in a heatmap.

B. Modeling accuracy of models involving either all features or individual features. Each column corresponds to the feature(s) (experiments and stages) involved in constructing statistical models for either the binding peaks of the transcription factors or the HOT regions (represented by the rows).

Fig. S39: Machine learning procedure for modeling TF binding peaks

The *C. elegans* genome was divided into bins of 100 bp. Histone methylation and binding signals of RNA Pol II were used as features to distinguish bins which intersect with the binding peaks from those which do not, using the machine learning method support vector machines (SVMs). Models were learned from the training portion of the data sets and evaluated on a separate testing portion.

Fig. S40: TF binding model accuracy

A. Modeling accuracy of integrative models. Each curve represents the accuracy of an integrative model involving all features together used to predict either the binding peaks from a TF binding experiment or HOT regions from the genomic background. See the caption for Fig. S39B for the learning procedure. The accuracy of the models is represented here by receiver-operator-characteristic (ROC) curves.

The whole genome was divided into bins of 100bp in size. Bins within the binding peaks of specific TF (or within the HOT regions) were defined as the positive examples, and an equal number of other bins were randomly sampled from the whole genome as the negative examples. These examples were used to train and test machine learning models using cross-validation. The number in each cell corresponds to the accuracy of the model, measured by the area under the receiver-operator-characteristic (ROC) curve, AUROC. The receiver operator characteristic is a plot of true positive rate against false positive rate for a set of ranked predictions. If all the ground truth positives are ranked higher than the ground true negatives, the curve goes from the origin vertically up to the point (0, 1), and then horizontally to (1, 1). In this case, the area under the curve has the maximum value of 1. If all the ground truth negatives are ranked higher than the ground true positives, the area under the curve has the minimum value of 0. A random ranking has an expected area under the curve of 0.5. In general, a larger area under the curve indicates a higher consistency between the predictions and the ground truth.

B. Distinguishing binding peaks of different TFs. Each bar shows the accuracy with which a model distinguishes the binding bins of a TF experiment from random binding bins of other TF experiments (instead of the genomic background as in part A). The last column shows that the HOT regions can be accurately separated from other TF binding sites using the chromatin features.

Fig. S41:Average signals of some chromatin marks at the binding-peaks and non-binding-peaks of TFs

Each panel shows the average signal of a chromatin mark at the binding-peaks and non-binding-peaks of each TF-binding dataset. H3K4me2 and H3K4me3 are enriched in TF-binding peaks as compared to non-peaks, while H3K9me3 are depleted. Among the TFs, H3K4me2 and H3K4me3 are in general more enriched at the peaks of CEH-14, CEH-30, LIN-13, LIN-15B and MEP-1 as compared to the other TFs.

Fig. S42: Developmental stage-specific models

The accuracy of models specific for individual developmental stages (involving predictors only from that stage) are shown. For each TF, the heights of the three bars correspond to the accuracies of the models for distinguishing binding peaks of different TFs, involving predictors measured in (from left to right) embryos only, L3 only, and both stages. Notice that the results for the last case were also shown in Fig. S40.)

Fig. S43: Combination of chromatin and sequence features

Potential binding sites of HLH-1 were identified by using two known sequence motifs in Jaspar (*108*). Chromatin features were used to model general binding active regions (BAR+) which are not specific to any DNA-binding proteins. The prediction model assigns a probability value for each region to indicate its likelihood of being in BAR+. By varying the probability threshold, different sets of BAR+ regions were identified. At each threshold, three sets of regions were compared: all general binding active regions (BAR+), all regions with high motif scores (PWM+), and binding active regions with high motif scores (BAR+PWM+).PPV represents the fraction of true positives in all positive predictions by the model.

Fig. S44: Coverage of evolutionarily constrained regions by genomic features

From the six-way alignment of *C. elegans*, *C. briggsae, C. brenneri, C. japonica*, and *P. pacificus,* we identified the portion of the genome under evolutionary constraint as described in the SOM. From this, we calculated the overlap with pre- and post-modENCODE functional elements in order to determine the proportion of constrained regions that can be explained by known classes of functional elements. Bars indicate the coverage of the whole and constrained genome by individual element classes with darker colors indicating the proportion of the class that overlaps with previous classes and lighter colors indicating the proportion uniquely contributing to coverage. The lines indicate cumulative coverage. Red and orange bars and lines correspond to coverage of the entire genome, while blue and light blue corresponds to coverage of the constrained 29.6% of the genome.

Element sets are as follows: *WB CDS:* WormBase coding exons from release WS190 that are completely confirmed by cDNA and EST evidence (see Table S14); *WB 5' UTR, WB 3' UTR:* WormBase UTRs from the same release; *WB partially confirmed CDS:* WormBase coding exons that are partially confirmed by overlapping cDNA and ESTs; *ME CDS:* coding exons that are fully supported by transcriptome sequencing data generated by modENCODE; *ME 5' UTRs, ME 3' UTRs:* UTRs that are supported by modENCODE transcriptome sequencing; *WB predicted CDS:* unconfirmed coding exons from WormBase called by *ab initio* gene prediction algorithms (added to the cumulative plot after the confirmed gene elements to show the small additional contribution that predicted exons make to coverage); *ncRNA*: modENCODE nonfadingRNA annotations; *Pseudogene*: modENCODE pseudogene annotations; *TF binding sites*: modENCODE binding sites for 23 transcription factors; *Chromatin associated proteins*: the union of modENCODE binding site peaks for the chromatin associated proteins HCP-3, LEM-2, MES-4 and HRG-1; *Dosage compensation factors*: the union of modENCODE binding site peaks for DPY-27, DPY-28, MIX-1, SDC-2 and SDC-3.

Fig. S45: Conservation enrichment analysis

Expanded version of Fig. 8 comparing spectral enrichment and overall enrichment of 12 datasets.

Datasets for A "CDS, UTR and ncRNA regions". *ncRNA:* non coding RNAs identified by modENCODE; *miRNA:* microRNAs identified by modENCODE; *5' UTR, 3' UTR:* WormBase 5'- and 3'-UTRs confirmed by EST alignments; *CDS:* All modENCODE validated coding regions.

Datasets for B "chromatin interacting protein sites". *Dosage compensation:* the union of binding site peaks for the factors DPY-27, DPY-28, MIX-1, SDC-2 and SDC-3. *TF binding sites:* Transcription factor (TF) binding sites identified by modENCODE. The union of the binding sites from the following experiments were counted as transcription factor binding sites: ALR1 L2, BLMP L1, CEH14 L2, CEH30 LE, EGL5 L3, EGL27 L1, ELT3 L1, EOR1 L3, GEI11 L4, HLH1 EMB, LIN11 L2, LIN13 EMB, LIN15B L3, LIN39 L3, MAB5 L3, MDL1 L1, MEP1 EMB, PES1 L4, PHA4 EMB, PHA4 L1, PHA4 L2, PHA4 LE, PHA4 stvL1, PHA4 YA, PQM1 L3, SKN1 L1, and UNC130 L1. *HOT:* ChIP target regions occupied by at least 15 TFs. The most stringent definition of HOT region was applied here, counting only those bases with at least 15x overlap from peaks from TF group, while for PHA4 only peaks from L2 were considered. *Remaining chromatin interacting protein sites:* the union of binding site peaks for the factors HCP-3, LEM-2, MES-4 and MRG-1 .

Datasets for C "introns, pseudogenes and unannotated regions". *pseudogenes:* modENCODE pseudogenes annotations; *introns:* modENCODE intron annotations. *Unannoated regions*: All genomic regions not covered by any of the preceding datasets from A, B, or C.

Spectral plots show enrichment relative to constrained bases in non-CDS portion of the genome. Points above and below the dotted horizontal line are enriched and depleted, respectively, relative to expectation drawn from a random distribution of similar size fragments from the non-CDS portion of the genome. The inset column bars show overall enrichment relative to entire genome. Only peaks that did not overlap CDS, UTR and ncRNA regions were considered, so as to remove any artificial conservation artifact due to the background conservation of such elements. For the analysis of introns we similarly excluded introns that overlap a CDS or UTR on the opposite strand.

Fig. S46: PhastCons score correlation with peak centers in modENCODE peak calls

Aggregate conservation scores for peaks from three representative transcription factors, LIN-15B, HLH-1 and ALR-1. PhastCons conservation scores increase towards the center of called peaks and reach their maximum near peak centers identified by PeakRanger.

Fig. S47: Saturation of TF binding

Saturation of the binding sites of 23 *C. elegans* transcription factors (including 6 stages for PHA-4) over the WS190 genome with coding sequence bases and Pol II binding site bases removed. No more than 5% of the bases are covered by these factors. These experiments include: ALR-1 L2, BLMP-1 L1, CEH-14 L2, CEH-30 LE, DPY2-7 EMB, EGL-5 L3, EGL-27 L1, ELT-3 L1, EOR-1 L3, GEI-11 L4, HLH-1 EMB, LIN-11 L2, LIN-13 EMB, LIN-15B L3, LIN-39 L3, MAB-5 L3, MDL-1 L1, MEP-1 EMB, PES-1 L4, PHA-4 EMB, PHA-4 L1, PHA-4 L2, PHA-4 LE, PHA-4 stvL1, PHA-4 YA, PQM-1 L3, SKN-1 L1, and UNC-130 L1.

Fig. S48: Comparison of coverage between ENCODE pilot and modENCODE *C. elegans* project

The upper panel illustrates the distribution of evolutionary constrained bases in the human ENCODE pilot, while the lower panel illustrates the distribution in modENCODE. The pie charts demonstrate the relative proportion of constrained and unconstrained bases according to the definitions used in this paper and the ENCODE pilot, while the stacked column chart shows the coverage of the constrained bases among various classes of annotation. While the number of bases annotated by the *C. elegans* modENCODE project is considerably higher than the ENCODE project, owing to the fact that the modENCODE projects target whole organism genomes while the ENCODE pilot focused on 1% of the human genome, the percentage of experimental annotations added by each are very similar. The percentage of regions that remain unannotated is much smaller in *C. elegans* owing primarily to the fact that more dense coding regions were found in worm relative to human.

Fig. S49: RNAPII ChIP-seq signal aggregation in human and *C. elegans*

Aggregation of RNA polymerase II (RNAPII) ChIP-seq signal over TSSs for *H. sapiens* (HeLa cells, (*61*)) and C. elegans (L4). Annotation sources used were CCDS genes (*H. sapiens*, (*96*)) and wormbase build ws180 (*C. elegans*, (*97*)). Signals were scaled to share the same maximum height and show that RNAPII signals exhibit similar profiles between organisms.

Fig. S50: Comparison of histone marks in *C. elegans* early embryos and human CD4T cells

Average gene profiles around the TSS and TTS of RefSeq genes are shown for human CD4+T cell ChIP-seq data of various histone marks (*98*). The top 20% expressed genes are shown in dark red, the bottom 20% genes are shown in light red. The ChIP-seq data were processed as described in (*99*)

Supplementary Tables

Table S1a: Data overview for RNA sequencing and expression tiling arrays

modENCODE Transcriptome Experiments							
Stage/Condition	**PolyA selected RNA-seq**		**Small RNA-selected RNA-seq**		**3' UTR-selected RNA-seq**		**Expression Tiling Arrays**
	Substages / Subconditions	Reads (Millions)	Substages / Subconditions	Reads (Millions)	Substages / Subconditions	Reads (Millions)	Substages / Subconditions
Embryo	2	139	8	51	1	0.5	3
L1	2	111	1	10	1	0.2	1
L2	1	33	1	10	1	0.4	1
L3	1	29	1	9	1	0.2	1
L4	2	78	1	9	1	0.3	3
Adult herm.	4	195	10	106	1	0.1	4
Male	2	60	1	12	1	0.2	1
Dauer	3	89			4	0.2	
Mixed stage			1	5	3	11	
Isolated Tissues			4	16			26
Infected Organisms	2	97					4
Total	19	831	28	228	14	13.1	44

This is an overview of the raw experimental data present in the February 2010 data freeze from the transcription analysis portion of the project. Developmental substages, isolated tissues, and several mutant strains have been collapsed into single columns; the counts in each "Substages/Subconditions" column give the numbers of substages, tissues and/or mutant strains examined. The background strain is N2, unless otherwise noted. Substages include: Embryo: early embryo, late embryo, mixed-stage embryo, one-cell stage embryo, post-gastrulation embryo, two-to-four cell embryo; L1: N2, *lin-35*; L4: hermaphrodite, JK1107 soma, L3-L4; Dauer: *daf-2* dauer larva (entry, mid, exit), *daf-3, daf-7, daf-9, daf-11*; *Adult hermaphrodite*: adult (includes controls for pathogen assays), young adult, *spe-9* adult (0, 5, 8, 12 days), JK1107 soma, L4-YA; *Male*: *him-8* embryo, *dpy28(y1);him-8(e1489)* L4 male, *him-8* adult male; Isolated tissues: GABA neurons, A-class motor neurons, AVA neurons, body wall muscle, coelomocytes, dopaminergic neurons, GABA motor neurons, germline precursor, hypodermal cells, intestine, panneural, BAG neurons, pharyngeal muscle, PVC neurons, excretory cell, glutamate receptor neurons, PVD & OLL neurons, cephalic sheath cells (CEPsh), spermatids, oocytes, gonad; *Infected Organisms (3 pathogens)*: *E. faecalis, P. luminscens, S. marcescens.* "PolyA selected RNA-seq" refers to RNA sequencing of polyA-selected libraries. "3' UTR-selected RNA-seq" refers to 3'-RACE and other experimental strategies designed to sequence the 3' ends of transcribed genes.

Table S1b: ChIP-chip, ChIP-seq, and other chromatin-characterization experiments

Transcription Factor ChIP-Seq

Transcription Factor	Early Embryo	Mixed Embryo	Late Embryo	L1	L2	L3	L4	Young Adult	Other
HLH-1		X							
LIN-13		X							
MEP-1		X							
PHA-4		X	X	X	X			X	Starved L1
CEH-30			X						
BLMP-1				X					
EGL-27				X					
ELT-3				X					
MDL-1				X					
SKN-1				X					
UNC-130				X					
ALR-1					X				
CEH-14					X				
LIN-11					X				
EGL-5						X			
EOR-1						X			
LIN-15B						X			
LIN-39						X			
MAB-5						X			
PQM-1						X			
GEI-11							X		
PES-1							X		
POL-II	X		X	X	X	X	X	X	

Chromatin Modification ChIP-chip

Modification	Early Embryo	Mixed Embryo	Late Embryo	L1	L2	L3	L4	Young Adult	Other
H3K27Ac	X					X			
H3K27me3						X			
H3K36me1	X					X			
H3K36me2	X					X			
H3K36me3	X					X			
H3K4me1	X					X			
H3K4me2	X					X			
H3K4me3	X					X			
H3K79me1	X					X			
H3K79me2	X					X			

H3K79me3	**X**					**X**			
H3K9Ac	**X**					**X**			
H3K9me1	**X**					**X**			
H3K9me2	**X**					**X**			
H3K9me3	**X**					**X**			
H3K20me1	**X**					**X**			
H4tetraAc	**X**								
H4K8Ac						**X**			

Chromatin-Associated Proteins ChIP-chip

Factor	**Early Embryo**	**Mixed Embryo**	**Late Embryo**	**L1**	**L2**	**L3**	**L4**	**Young Adult**	**Other**
CBP-1		**X**							
DPY-26		**X**							
DPY-27	**X**	**X**					**X**		
DPY-28		**X**							
H3	**X**								
H4						**X**			
HCP-3	**X**	**X**							
HTZ-1		**X**							
LEM-2		**X**							
MES-4	**X**								
MIX-1		**X**							
MRG-1	**X**								
NPP-13		**X**							
POL-II	**X**	**X**					**X**		
SDC-2		**X**							
SDC-3		**X**							
Chromatin salt fractionation		**X**							
Nucleosomes (MNase-seq)	**X**	**X**						**X**	**glp-1 adults, fem-2 adults**

This is an overview of the raw experimental data present in the February 2010 data freeze from the transcription factor and chromatin-structure aspects of the project. All transcription factors were analyzed in replicate by ChIP-seq. All chromatin modifications and chromatin-associated proteins were analyzed by ChIP-chip, with the exception of DPY-27, which was analyzed by ChIP-seq as well as ChIP-chip. All experiments were performed on N2, unless noted in the "Other" column.

Table S1c: Inferred genomic elements

Inferred Genomic Elements

Element	Representative Developmental Stages					
	Embryo	L1	L2	L3	L4	Young Adult
Coding transcripts	35,097	35,568	32,027	34,216	34,471	33,999
TSSs	14,854	16,257	14,411	14,004	10,949	13,960
TARs	39,328	42,421	41,791	41,734	43,380	40,624
miRNAs	152	127	126	130	133	133
other ncRNAs	895	936	981	781	823	859
TF Peaks (# factors)	17,147 (5)	26,944 (7)	8,060 (4)	16,149 (6)	3,749 (2)	551 (1)

Here we summarize genomic elements that have been inferred for each major element type across the developmental series. For simplicity, we have chosen a single representative subcondition for each stage. *Embryo:* early N2 embryo for all experiments except for the miRNA and other ncRNA experiments, which were performed on mixed embryonic stages from N2; *L1-L4:* L1 through L4 larva in the N2 strain; *YA:* Young adult N2 hermaphrodites

Table S2a: Sources of polyA sites in final integrated transcript set

Data source(s)*	proximal**	distal	total
3P-only	3570	4610	8180
3P+Mangone	2680	6291	8971
3P+Mangone+RNAseq	1676	4181	5857
3P+Mangone+RNAseq+Wb	273	534	807
3P+Mangone+Wb	97	358	455
3P+RNAseq	199	311	510
3P+RNAseq+Wb	13	22	35
3P+Wb	7	26	33
Mangone-only	3847	1794	5641
Mangone+RNA-seq	43	12	55
Mangone+Wb	10	22	32
RNA-seq-only	46	93	139
Wb-only	9	13	22

*3P = 3P-Seq from Jan et al. (*15*)
Mangone = Mangone et al. (*17*)
RNA-seq = those polyAs identified by this project
Wb = WormBase (WB170)

**proximal = a polyA site in a terminal exon that is not the most distal polyA site in the exon

Table S2b: *C. elegans* genes not identified as "transcribed" in 19 polyA RNA-seq samples

Type	Genome total	Not covered	% not found
nuclear hormone receptors	85	21	24.7
7TM/G-protein coupled receptor	1454	323	22.2
math-(meprin-associated Traf homology)	62	9	14.5
F-box	238	12	5
Zinc Finger	236	4	1.7

Stages and strains of worm RNA (polyA) sequenced include: embryonic *him-8*(e1489) (50% males), early embryos, late embryos, L1 *lin-35*(n1745), L1, L2, L3 dauer entry *daf-2*(e1370), dauer daf-2(e1370), dauer exit daf-2(e1370), L4, L4 males, JK1107 L4 (no gonad) *glp-1*(q224), young adults, aged adults (spe-9(hc88)), adults exposed to Harposporium spp (tentative assignment), and adults exposed to *S. marcescens.*

Table S3: Developmental stages and tissue samples of small RNA-seq and tiling array experiments.

RNA-seq and Tiling Array Samples	RNA-seq abbreviations
Developmental stages for small RNA-seq experiments	
Young adult males (23dC)	
Mixed Embryo	
mid-L1 20dC 4hrs post-L1 stage larvae	
mid-L2 20dC 14hrs post-L1 stage larvae	
mid-L3 20dC 25 hrs post-L1 stage larvae	
mid-L4 20dC 36hrs post-L1 stage larvae	
Young adult 20dC 48hrs post-L1 stage larvae	
Young adult (23dC DAY 0 post-L4 molt)	
Adult 23dC 12 days post-L4 stage larvae	
Adult 23dC 5 days post-L4 stage larvae	
Adult *spe-9*(hc88) 23dC 8 days post-L4 molt	
Specific tissues for tiling array experiments	
embryo A-class motor neurons	
embryo AVA neurons	
embryo body wall muscle (v2)	
embryo coelomocytes	
embryo dopaminergic neurons	
embryo GABA motor neurons	
embryo germline precursor cells	
embryo hypodermal cells	
embryo intestine	
embryo panneural	
L2 A-class neuron	
L2 body wall muscle	
L2 coelomocytes	
L2 excretory cell	
L2 GABA neurons	
L2 glutamate receptor expressing neurons	
L2 intestine	
L2 panneural	
L3-L4 dopaminergic neuron	
L3-L4 hypodermal cells	
L3-L4 PVD & OLL neurons	
Young Adult Cephalic sheath (CEPsh)	
embryo BAG neurons*	
embryo PVC neurons*	
embryo pharyngeal muscle*	
Controls for tissue specific tiling array experiments	
embryo all cells reference	
L2 reference (mockIP)	
L3-L4 reference (mockIP)	

Young Adult reference (mockIP)	
Developmental stages of tiling array experiments	
early embryo 20dC 0-4hrs post-fertilization*	EE
late embryo 20dC 6-12hrs post-fertilization N2	LE
L2 polyA enriched 20dC 14hrs post-L1 N2	
L1 20dC 0hrs post-L1 N2	L1
L2 25dC 14hrs post-L1 N2	L2
L3 25dC 25hrs post-L1 N2	L3
L4 25dC 36hrs post-L1 N2	L4
young adult 25dC 42hrs post-L1 N2	YA
male L4 25dC 36hrs post-L1 CB4689	L4 male
gonad from young adult 20dC 42hrs post-L1 N2	
soma-only mid-L4 25dC 36hrs post-L1 JK1107	L4 soma
pathogen *S marcescens* 25dC 24hr exposure post-adulthood	Sm
pathogen *S marcescens* 25dC 48hr exposure post-adulthood	
pathogen *E faecalis* 25dC 24hr exposure post-adulthood	
non-pathogen control 25dC 24hr exposure post-adulthood	Sm ctrl
non-pathogen control 25dC 48hr exposure post-adulthood	
pathogen *P luminscens* 25dC 24hr exposure post-adulthood	
Other samples not included above used for RNA-seq experiments	
mixed embryo (*him-8*)	MxE
L1 (*lin-35)*(n745)	L1 *(lin-35)*
dauer entry *daf-2* 25dC 48hrs post-L1	dauer entry
dauer *daf-2* 25dC 91hrs post-L1	dauer
dauer exit *daf-2* 25dC 91hrs 15dC 12hrs post-L1	dauer exit
Aged adult (*spe-9*) 23dC 8 days post-L4 molt	aged adult
Adult *Harposporium spp* control E. coli OP50 exposed 2 24hrs	Hs ctrl
Adult *Harposporium spp* exposed 2 24hrs	Hs

* Four samples were not included in ncRNA prediction and further analysis of ncRNAs because they were released after the ncRNA companion paper was submitted.

Table S4: Summary of cell and stage specific tiling array results

Feature class	FDR	Samples	# of features WS199[1]	% of features WS199[2]
Annotated exons (unique) of coding genes overlapping with nrTARs[3]		cells & stages	119,521 exons	87.1% (137,193)
		cells	116,929 exons	
		stages	100,658 exons	
Annotated coding genes with exons overlapping with nrTARs		cells & stages	18,183 genes	91.3% (19,912)
		cells	18,049 genes	
		stages	15,400 genes	
Exons of integrated transcript models (unique) overlapping with nrTARs		cells & stages	138,433 exons	87.8% (157,612)
		cells	135,654 exons	
		stages	116,799 exons	
Integrated transcript models with exons overlapping with nrTARs		cells & stages	19,325 genes	88.8% (21,774)
		cells	19,173 genes	
		stages	16,152 genes	
Gene models detected[4]	5%	cells & stages	17,452 genes	87.7% (19,912)
	5%	cells	17,075 genes	
	5%	stages	15,822 genes	
Gene models detected (FDR-corrected)[5]	0.14%	cells & stages	14,279 genes	71.7% (19,912)
	0.17%	cells	13,149 genes	
	0.71%	stages	13,713 genes	
Gene models differentially expressed (at least 2 fold)[6]	5%	cells & stages	13,320 genes	66.9 % (19,912)
	5%	cells	10,598 genes	
	5%	stages	9,552 genes	
Gene models differentially expressed (at least 2 fold) (FDR-corrected)[7]	0.11%	cells & stages	11,299 genes	56.7 % (19,912)
	0.20%	cells	7,983 genes	
	0.24%	stages	8,606 genes	

[1] Protein-coding gene models are as described in WS199. Overlapping features were merged to produce a total of 19,912 gene models.

[2] Experimental results were calculated for 19,181 genes on the Affymetrix *C. elegans* 1.0R Tiling Array with ≥ 3 nonrepetitve exon probes. % of features is based on the total # of genes in WS199 (19,912) which is substantially similar to WS190 (20,121).

[3] non-redundant Transcriptionally Active Regions (nrTARs): Contiguous stretch of nucleotides all of which are inclusive to a TAR detected in ≥ 1 of the samples.

[4] The False Discovery Rate (FDR) of 5% was calculated for each sample independently and the total number of genes tabulated from the union of these results.

[5] Correction for potential accumulation of false positives arising from multiple testing. The FDR of each sample (5%) was divided by the cumulative number of samples for each category considered: cells & stages = 37; cells = 30; stages = 7.

[6] The False Discovery Rate (FDR) of 5% was calculated for each independent comparison and the total number of genes tabulated from the union of these results.

[7] Correction for accumulation of false positives arising from multiple testing. The FDR of each sample (5%) was divided by the cumulative number of comparisons for each category considered: cells & stages = 46; cells = 25; stages = 21. (see supplemental methods for this table)

Table S5: Different types of known ncRNAs

Type	**Number**
rRNA	19
scRNA	1
snRNA	94
snlRNA	4
snoRNA	139
tRNA*	630
miRNA	174
Total	**1061**
The miRNAs are collected from miRBase 14, and other ncRNAs are collected from WormBase 200. *24 tRNAs are from Mitochondria.	

Table S6: Annotated regions used for the training of machine learning methods (21K-set) – tiling array TARs

	Transcribed regions overlapped with confirmed annotations (Training Set)		
	CDS	UTR	known ncRNA[d]
	97.4%[a]	86.7%[a]	58.9%[a]
Total number of bases	5,117,511 (9,714,480)[b]	2,682,448 (7,498,856)[b]	51,928 (181,034)[b]
Total number of windows	51,721 (14,230)[b]	27,084 (9,854)[b]	489 (225)[b]
Number of windows with known 2" structure[c]	318 (183)[b]	320 (201)[b]	305 (160)[b]

[a] Fraction of annotated elements overlapped with tiling array TARs

[b] Values in the parenthesis are counted for the TARs, from which the fragmented windows are derived.

[c] Predicted with RNA secondary structure models from Rfam

[d] This is just the gold standard set and doesn't include any unconfirmed ones. Only 10% of known ncRNA were sampled because of large number of annotated tRNAs in the gold standard set.

Table S7: Performance of our integrated method (21K-set) on tiling array TARs[a] with three different ways to define element classes in the gold-standard set

Class definition 1		Class definition 2		Class definition 3	
Element class	AUC	Element class	AUC	Element class	AUC
ncRNA	0.9718	ncRNA	0.9246	ncRNA	0.9418
Coding exon	0.9718	Coding exon	0.7485	Coding exon	0.7361
		3’ UTR	0.7448	5’ and 3’ UTR	0.7315

[a] The minimum length of a TAR is 100nt. Large TARs are binned into 100nt windows with a step size of 75nt.

Table S8: Annotated and novel tiling array TARs going into 21K-set of ncRNAs

	Transcribed regions overlapped with annotated exons or known ncRNA (Confirmed and predicted)		Novel transcribed regions		
	Exon (81.3%)[a]	known ncRNA (13.1%)[a]	CDS-like[b]	UTR-like[b]	ncRNA-like[b]
Total number of bases	32,744,074 (33,532,732)[c]	265,250 (640,719)[c]	45,208 (134,041)[c]	368,771 (1,048,017)[c]	4,352,048 (6,503,326)[c]
Total number of windows	396,551 (77,131)[c]	2,547 (1,331)[c]	441 (194)[c]	3,294 (1,983)[c]	45,913 **(21,521)**[c]
Number of windows with known secondary structure)[d]	7,314 (3,988)[c]	961 (519)[c]	26 (19)[c]	152 (113)[c]	3,537 (2,083)[c]

[a] Fraction of annotated elements overlapped with tiling array TARs.

[b] In the prediction, if the probability of being ncRNA is larger than 0.009 but less than 0.297, it is ncRNA-like; if the probability of being a UTR is larger than 0.297 or less than 0.692, it is UTR-like; otherwise, if the probability of being CDS is larger than 0.692, it is CDS-like. The cut-offs are determined from the ROC curves.

[c] The long TARs are fragmented into small windows, and values in the parenthesis are counted for the original TARs.

[d] Secondary structure is predicted from Rfam/INFERNAL.

Table S9: Co-expression clusters of coding transcripts and novel ncRNA candidates (7K-set)

Cluster	Coding Transcripts	Candidate ncRNA bins*	Total	Candidate ncRNA bins %	Coding Transcripts GO enrichment**	Coding Transcripts GO depletion**
0	1199	1193	2392	49.87%	Transmembrane proteins, Receptors, Signal transducer activity	Protein binding, Development, Growth regulation
1	2727	4003	6730	59.48%	Chromatin assembly, DNA binding, Organelle organization	Protein binding
2	4210	604	4814	12.55%	Ion channel, Receptor, Membrane, Signal transducer activity, Transcription	Protein binding, Laval development, Growth regulation, Organelle, Cell cycle
3	4377	424	4801	8.83%	Development, Growth regulation, Reproduction	Receptor, Signal transducer activity
4	4274	356	4630	7.69%	Development, Cell cycle, Growth regulation, Reproduction	Receptor, Membrane, Signal transducer activity, Transcription factor activity
5	2931	362	3293	10.99%	Lipid metabolism, Sugar binding, Anion transport	Development, Receptor, Signal transducer activity, Growth regulation
6	577	1805	2382	75.78%	No significant enrichment	No significant depletion
7	2185	115	2300	5.00%	Membrane, Signal transducer activity, Receptor	Development, Growth regulation, Organelle
8	1548	149	1697	8.78%	Metabolism, Kinase	Development, Signal transducer activity, Receptor, Expression regulation
9	276	1276	1552	82.22%	No significant enrichment	No significant depletion
10	1182	75	1257	5.97%	Receptor, Membrane, Signal transducer activity, Ion	Development, Growth regulation, Reproduction

					channel	
11	677	527	1204	43.77%	Transcription factor activity, Ion binding,	No significant depletion
12	660	42	702	5.98%	Chromatin assembly, DNA binding, Organelle organization	Membrane
13	499	63	562	11.21%	Organelle	No significant depletion
Total	27322	10994*	38316			

The total RNA tiling array data at different tissues and stages were used to calculate the expression level of coding transcripts and candidate ncRNA bins.
*The candidate ncRNA bins are merged into 7k-set novel ncRNA candidates.
** The cut-off of enrichment or depletion is p value <0.01.

Table S10: Total mapped reads, numbers of peaks bound by each of 23 factors (22 TFs and one dosage compensation factor, 28 experiments in total) from ChIP-seq.

	# of Binding Sites	# of Total Mapped Reads	
	Narrow Peaks[1]	**GFP**	**Input**
ALR-1 L2	2383	3,746,542	2,506,542
BLMP-1 L1	6833	13,699,035	7,832,710
CEH-14 L2	1467	4,369,374	1,124,270
CEH-30 LE	1605	6,915,024	6,288,570
EGL-27 L1	135	3,402,816	2,862,812
EGL-5 L3	975	2,970,537	1,861,526
ELT-3 L1	1970	5,558,439	6,612,443
EOR-1 L3	3197	2,386,942	3,327,484
GEI1-1 L4	4356	2,744,559	4,498,845
HLH-1 MxE	512	4,052,296	2,488,302
LIN1-1 L2	1352	2,942,539	4,563,448
LIN1-3 MxE	680	5,108,200	9,056,899
LIN-15B L3	1726	2,024,367	6,045,335
LIN-39 L3	2954	3,399,898	1,993,494
MAB-5 L3	3763	3,517,148	3,568,848
MDL-1 L1	1691	4,134,371	4,264,998
MEP-1 MxE	5333	4,239,180	5,082,534
PES-1 L4	3088	3,417,784	2,630,081
PHA-4 MxE	3786	7,719,682	9,994,939
PHA-4 L1	4648	15,222,883	11,556,011
PHA-4 L2	6203	4,593,131	2,284,558
PHA-4 LE	4569	4,574,629	6,295,331
PHA-4 StvL1	5792	17,845,198	26,819,222
PHA-4 YA	3568	5,123,545	10,555,219
PQM-1 L3	954	2,626,971	6,505,184
SKN-1 L1	3279	4,517,511	2,474,805
UNC-130 L1	3401	3,174,776	5,312,775
DPY-27 MxE[2]	135	2,074,238	7,578,449

[1] Narrow binding peaks defined by PeakRanger.
[2] Dosage compensation factor.
MxE: mixed embryo; LE: late embryo; StvL1: starved L1 YA: young adult

Table S11: GO analysis of genes associated with HOT regions

GO ID	Name	P-value	Sample frequency	Background frequency	Genes
0040007	growth	1.49E-19	82/153 (53.6%)	2845/15340 (18.5%)	*rps-22, rps-12,* Y87G2A.1, *hsp-1, rps-25, lin-54, rps-28,* F36A2.7, *rps-24, pfd-1, dpy-23,* Y82E9BR.3, *glit-1, vps-32.1, xbx-5, sys-1, rpl-5, wrt-5, rpl-43,* R11D1.9, Y65B4BR.5, K12H4.5, Y49A3A.1, Y71H2AM.5, *rps-1,* K10D2.5, *puf-9, taf-4, rpl-3, hsr-9, eif-3.B, epc-1,* F17C11.9, W04A4.5, *mei-2,* K04G7.1, *rps-30, ash-2, wip-1,* H28O16.1, Y48A6B.3, ZK550.3, *cco-2, rpl-13, his-37, mbk-1, set-16, vha-8, kbp-4, cap-2, nipi-3,* C34C12.2, F48C1.4, *pbs-2, sor-1, dpm-3,* T23F11.1, *eft-4, rpl-6, rpl-7, ekl-4, rpl-32, rpl-22,* E02D9.1, *emo-1, atad-3, nuo-1,* LLC1.3, *cct-1,* Y51H4A.15, *cco-1, rpn-3, rps-26, rpl-24.1, rpl-14, prp-8, mdt-19, rpl-35, rfc-4, mdt-26, htz-1, eft-2*
0009792	embryo development ending in birth or egg hatching	1.67E-17	82/153 (53.6%)	3054/15340 (19.9%)	*rps-22, rps-12, cbp-1, atx-2, hsp-1, rps-25, lin-54,* F36A2.7, F40F11.2, *pfd-1, dpy-23,* Y82E9BR.3, *vps-32.1, sys-1, rpl-5, rpl-43,* R11D1.9, Y65B4BR.5, K12H4.5, Y49A3A.1, T08B2.11, Y71H2AM.5, *rps-1, taf-4, rpl-3, eif-3.B, epc-1,* F17C11.9, W04A4.5, *mei-2,* K04G7.1, *ile-2, daf-21, wwp-1, ash-2, wip-1,* H28O16.1, *hsp-60,* ZK550.3, *klc-1, mdl-1, vig-1, cco-2, rpl-13, his-37, pqn-51, set-16, cls-2, tre-1, vha-8, kbp-4, cap-2,* F25E2.2, *cpt-2, nipi-3,* F48C1.4, *let-268, pbs-2, eft-4, rpl-6, rpl-7, ekl-4, rpl-22, dnj-11, emo-1, atad-3, nuo-1,* LLC1.3, *cct-1,* Y51H4A.15, *cco-1, rpn-3, rps-26, rpl-24.1, rpl-14, prp-8, mdt-19, rpl-35, rfc-4, mdt-26, htz-1, eft-2*
0005737	cytoplasm	7.43E-15	47/153 (30.7%)	1130/15340 (7.4%)	*rps-22, rps-12, cbp-1, atx-2, hsp-1, egl-30, trap-3, rps-28, rps-24, ddp-1, pfd-1, dpy-23, sys-1, rpl-5, rpl-43,* R11D1.9, Y71H2AM.5, *rps-1, puf-9, rpl-3, ain-1, F17C11.9, rps-30, daf-21, wwp-1, hsp-60, eat-16, deb-1, rpl-13, cls-2, vha-8, tra-4, cap-2, unc-108, let-268, eft-4, rpl-6, rpl-7, rpl-32, rpl-22, nuo-1,* LLC1.3, *cco-1, rps-26, rpl-24.1, rpl-14, rpl-35*
0005840	ribosome	1.65E-13	19/153 (12.4%)	141/15340 (0.9%)	*rps-22, rps-12, rps-28, rps-24, rpl-5, rpl-43,* R11D1.9, *rps-1, rpl-3, rps-30, rpl-13, rpl-6, rpl-7, rpl-32, rpl-22, rps-26, rpl-24.1, rpl-14, rpl-35*

Table S12: Expression correlation of transcription factors with target genes and non-target genes.

TF	Target	non-Target	Z-score	P-value
ALR-1	-0.007457	-0.008889	0.129717	0.896799
BLMP-1	-0.066699	-0.041505	-2.489284	0.012836
CEH-14	-0.096922	0.006595	-11.395537	0
EGL-27	0.128903	-0.034187	12.274861	0
EGL-5	0.0441	0.007088	5.185788	0
ELT-3	0.014695	-0.019066	3.251677	0.001165
EOR-1	0.133573	-0.042516	21.571089	0
GEI-11	0.068925	-0.005121	2.314903	0.021352
LIN-11	-0.081399	-0.048996	-2.416469	0.015807
LIN-15B	0.132488	-0.047239	21.460631	0
LIN-39	0.072205	-0.013786	11.372508	0
MAB-5	0.026751	0.009732	1.416596	0.156752
MDL-1	-0.0611	0.037063	-17.003577	0
PES-1	0.195083	-0.037515	20.789856	0
PHA-4	0.016003	-0.052655	16.739494	0
PQM-1	0.312876	0.016214	32.681676	0
SKN-1	0.090011	-0.017028	29.322469	0
UNC-130	-0.080614	-0.028788	-1.698375	0.090283

For each TF, the Pearson correlation coefficients of the expression level of the TF with those of its target genes and non-target genes were calculated across the 7 developmental stage time course. The significance of difference between target and non-target genes was calculated using t-test.

Table S13: Overview of PicTar-predicted miRNA target sites within 3'UTRs of the aggregated integrated transcript set (see text for details).

	3 species conservation	**5 species conservation**
Number of miRNAs analyzed	183	183
Number of 3'UTRs analyzed	25,539	25,539
Number of genes analyzed	14,519	14,519
Number of target sites detected	20,427	8,810
Number of 3'UTRs with target sites	4,866	2,406
Number of genes with target sites	2,349	1,162
Number of miRNAs that target a 3'UTR	182	178

Table S14: Overlap of 4.1 Mb of residual constrained blocks with various genomic elements.

Genomic elements	observed base pair overlap	expected by GSC simulation	Ratio obs/exp	p-value
Introns	0.47	0.35	1.3	1e-34
Intra-genic regions	0.27	0.18	1.5	1e-34
1000 bp upstream of gene TSS	0.19	0.18	1.06	2.7e-7
1000 bp downstream genic regions	0.18	0.23	0.78	1e-34

Table S15: This table shows a comparison of the amount of binding for a few selected transcription factors between *C.elegans* as compared to those from the ENCODE pilot project for human.

Pilot ENCODE Human TF Binding	Total Binding (bp)	Percentage of Pilot Regions (29.96Mb) Bound	Intergenic Binding (bp)	Percentage of Intergenic Pilot Regions (13.15Mb) Bound
STAT1	200,136	0.67%	175,890	1.34%
cFos	256,342	0.86%	203,557	1.55%
cJun	257,928	0.86%	211,855	1.61%
CTCF (32h)	281,292	0.94%	267,107	2.03%
CEBPε (32h)	346,901	1.16%	323,879	2.46%
Average	268,520	0.90%	236,458	1.80%
modENCODE *C.elegans* TF Binding	**Total Binding (bp)**	**Percentage of Genome Bound**	**Intergenic Binding (bp)**	**Percentage of Intergenic Regions (40.89Mb) Bound**
CEH-14 (L2)	290,252	0.29%	225,713	0.55%
EGL-27 (L1)	193,359	0.19%	180,394	0.44%
MAB-5 (L3)	337,306	0.34%	311,347	0.76%
PES-1 (L4)	752,681	0.75%	694,621	1.70%
PHA-4 (EMB)	921,663	0.92%	851,259	2.08%
Average	499,052	0.50%	452,667	1.11%

Table S16: This table shows a comparison between the amount of transcription between the modENCODE project for *C.elegans* and human from the pilot ENCODE project. The amount of transcription is broken down by the genic and intergenic components using GENCODE and WormBase WS190 annotation.

Pilot ENCODE Human Transcription	**Pilot Regions Transcribed (bp)**	**Genic Regions Transcribed (bp)**	**Percentage Genic Transcription**	**Intergenic Regions Transcribed (bp)**	**Percentage Intergenic Transcription**
Placenta PolyA	484,629	407,959	84.2%	76,670	15.8%
HeLa PolyA	905,973	507,108	56.0%	398,865	44.0%
modENCODE *C.elegans* Transcription	**Genome transcribed (bp)**	**Genic Regions Transcribed (bp)**	**Percentage Genic Transcription**	**Intergenic Regions Transcribed (bp)**	**Percentage Intergenic Transcription**
L2 PolyA	20,421,924	17,244,358	84.2%	3,177,566	15.6%

Table S17: Table of Wormbase versions used (Note here "T" denotes table so, "T7" means "supplementary table S7".)

Figure	Figure Title	Wormbase Version	#Wormbase Live genes	#Wormbase CDS (inc alt splice forms)
1	Transcriptome Features and Alternative Splicing	200, 170	39868, 23977	23973, 23224
2	Expression and Binding Dynamics	180	29500	23511
3	HOT Regions	190	29802	23771
4	Integrated Regulatory Network	170	23977	23224
5	Chromosome-scale domains of chromatin organization	170	23977	23224
6	Chromatin Patterns around Genes	170	23977	23224
7	Statistical Models Predicting Regulation and Expression from Chromatin Features	190	29802	23771
8	Relative proportion of annotations among constrained sequences	190	29802	23771
S1	ChIP-chip and ChIP-seq comparision	190	29802	23771
S2	Correlation of RNA expression levels for Young Adult between RNA-seq and tiling array platforms	190	29802	23771
S3	Numbers of RNA-seq Reads	170	23977	23224
S4	RNA sequencing depth analysis	170	23977	23224
S5	Transcript building	170	23977	23224
S6	A complex isoform example	170	23977	23224
S7	Features defined by RNAseq as compared to WormBase as of January, 2007 (WS170)	170	23977	23224
S8	Number of confirmed splice junction over time	170	23977	23224
S9	Proportion of splice junctions confirmed by various methods	170	23977	23224
S10	Saturation of discovery of additional ncRNAs and coding exons with additional RNA-seq data sets	190	29802	23771
S11	Number of stages and samples where a given gene or splice junction is observed	170	23977	23224
S12	Developmental stage-specific expression	190	29802	23771
S13	Lab batch effects	180	29500	23511
S14	Cumulative plot of isoform	190	29802	23771

	composition distribution			
S15	SOM clusters of transcripts with different developmental expression profiles	190	29802	23771
S16	Number of genes and transcripts shared between pairs of SOM clusters	190	29802	23771
S17	Classes of distinguishing features between isoforms with different developmental expression profiles based on SOM clustering	190	29802	23771
S18	Examples of read count distributions supporting differential expression of alternative transcript isoforms among developmental stages	180	29500	23511
S19	A breakdown on how the updated list of *C. elegans* pseudogenes was created	200	39868	23973
S20	Binning of long TARs built from tiling arrays	n/a	n/a	n/a
S21	Predicting ncRNAs	200, 170	39868, 23977	23973, 23224
S22	TF binding around non-coding RNAs	200, 170	39868, 23977	23973, 23224
S23	Co-occurrence of transcription factors	190	29802	23771
S24	Comparison of PeakSeq and PeakRanger peak calls	n/a	n/a	n/a
S25	Distribution of TF binding	190	29802	23771
S26	Control experiments for HOT regions	190	29802	23771
S27	HOT regions enrichments	190	29802	23771
S28	Higher gene expression level in HOT regions	190	29802	23771
S29	HOT regions are broadly expressed	190	29802	23771
S30	Pair-wise correlations of PHA-4 binding signal across different stages	190	29802	23771
S31	Examples of Pol II binding and expression	190	29802	23771
S32	Histone marks distribution over repetitive elements	190	29802	23771
S33	Promoters of chromosome X genes have higher GC content compared to autosomes	190	29802	23771
S34	Histone marks aggregation around	170	23977	23224

	TSS and TTS			
S35	TF sequence motif discovery	190	29802	23771
S36	TFs in the larval network	170	23977	23224
S37	Network motifs	170	23977	23224
S38	TF binding and chromatin features	182	29505	23523
S39	Machine learning procedure for modeling transcription factor binding peaks	n/a	n/a	n/a
S40	TF binding model accuracy	182	29505	23523
S41	Average signals of some chromatin marks at the binding-peaks and non-binding-peaks of TFs	182	29505	23523
S42	Developmental stage-specific models	182	29505	23523
S43	Combination of chromatin and sequence features	182	29505	23523
S44	Coverage of evolutionarily constrained regions by genomic features	190	29802	23771
S45	Conservation enrichment analysis	190	29802	23771
S46	PhastCons score correlation with peak centers in modENCODE peak calls	190	29802	23771
S47	Saturation of TF binding	190	29802	23771
S48	Comparison of coverage between ENCODE pilot and modENCODE C. elegans project	190	29802	23771
S49	RNAPII ChIP-seq signal aggregation in human and C. elegans	190	29802	23771
S50	Comparison of histone marks in C. elegans early embryos and human CD4T cells	190	29802	23771
T1a	Data overview for RNA sequencing and expression tiling arrays	190	29802	23771
T1b	ChIP-chip, ChIP-seq, and other chromatin-characterization experiments	190	29802	23771
T1c	Inferred genomic elements	170	23977	23224
T2a	Sources of polyA sites in final integrated transcript set	170	23977	23224
T2b	C. elegans genes not identified as transcribed in 19 polyA RNA-seq samples	170	23977	23224
T3	Develpmental stages and tissue samples of small RNA-seq and	190	29802	23771

	tiling array experiments			
T4	Summary of cell and stage specific tiling array results	199	39873	23973
T5	Different types of known ncRNAs	200	na	na
T6	Annotated regions used for the training of machine learning methods (21K-set) - tiling array TARs	170	23977	23224
T7	Performance of our integrated method (21K-set) on tiling array TARs[a] with three different ways to define element classes in the gold-standard set	170	23977	23224
T8	Annotated and novel tiling array TARs going into 21K-set of ncRNAs	170	23977	23224
T9	Co-expression clusters of coding transcripts and novel ncRNA candidates (7K-set)	170	23977	23224
T10	Total mapped reads, numbers of peaks bound by each of 23 factors (22 TFs and one dosage compensation factor, 28 experiments in total) from ChIP-seq	170	23977	23224
T11	GO analysis of genes associated with HOT regions	190	29802	23771
T12	Expression correlation of transcription factors with target genes and non-target genes	170	23977	23224
T13	Overview of PicTar-predicted miRNA target sites within 3'UTRs of the aggregated integrated transcript set (see text for details).	190	29802	23771
T14	Overlap of 4.1 Mb of residual constrained blocks with various genomic elements	190	29802	23771
T15	Sample comparison of C. elegans and human TF binding regions	190	29802	23771
T16	Sample comparison of C. elegans and human transcription	190	29802	23771

Supplementary References

1. modMine, http://intermine.modencode.org
2. UCSC Genome Browser, http://genome.ucsc.edu
3. Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo
4. Short Read Archive, http://www.ncbi.nlm.nih.gov/sra
5. WormBase, http://www.wormbase.org
6. L. R. Baugh, A. A. Hill, D. K. Slonim, E. L. Brown, C. P. Hunter, Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development* **130**, 889-900 (2003).
7. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733-739 (2010).
8. P. Chomczynski, N. Sacchi, Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction. *Anal Biochem* **162**, 156-159 (1987).
9. P. Kapranov *et al.*, Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919 (2002).
10. V. R. Iyer *et al.*, Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533-538 (2001).
11. A. Agarwal *et al.*, Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**, 383 (2010).
12. D. Kampa *et al.*, Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331-342 (2004).
13. T. E. Royce *et al.*, Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* **21**, 466-475 (2005).
14. L. W. Hillier *et al.*, Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res.* **19**, 657-666 (2009).
15. C. H. Jan, R. C. Friedman, J. G. Ruby, C. B. Burge, D. P. Bartel, Formation and regulation of 3′ untranslated regions in Caenorhabditis elegans, *Nature*, 10.1038/nature09616.
16. G. E. Merrihew *et al.*, Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res.* **18**, 1660-1669 (2008).
17. M. Mangone *et al.*, The Landscape of C. elegans 3'UTRs. *Science* **329**, 432-435 (2010).
18. RSEQtools, http://archive.gersteinlab.org/proj/rnaseq/rseqtools
19. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
20. T. L. Bailey, N. Williams, C. Misleh, W. W. Li, MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369-W373 (2006).
21. W. C. Spencer *et al.*, A spatial and temporal map of C. elegans gene expression. *Genome Res*., 10.1101/gr.114595.110.
22. M. Christensen *et al.*, A primary culture system for functional analysis of C-elegans neurons and muscle cells. *Neuron* **33**, 503-514 (2002).
23. R. M. Fox *et al.*, A gene expression fingerprint of C-elegans embryonic motor neurons. *BMC Genomics* **6**, 42 (2005).
24. S. E. Von Stetina *et al.*, Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C-elegans nervous system. *Genome Biol.* **8**, R135 (2007).
25. G. Zeller, S. R. Henz, S. Laubinger, D. Weigel, G. Ratsch, Transcript normalization and segmentation of tiling array data. *Pac Symp Biocomput*, 527-538 (2008).
26. S. Laubinger *et al.*, At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biol.* **9**, R112 (2008).
27. R. A. Irizarry *et al.*, Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, E15 (2003).
28. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

29. L. Gautier, L. Cope, B. M. Bolstad, R. A. Irizarry, affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315 (2004).

30. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).

31. C. E. Bonferroni, Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, 13–60 (1935).

32. G. K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, 3 (2004).

33. G. K. Smyth, in *Bioinformatics and Computational Biology Solutions using R and Bioconductor,* R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber, Eds. (Springer, New York, 2005), pp. 397-420.

34. IQseq, http://archive.gersteinlab.org/proj/rnaseq/IQSeq

35. documented source code for Deepseq9 algorithm at SourceForge, http://deepseq9.sourceforge.net

36. Supplemental files are available at http://www.modencode.org/publications/integrative_worm_2010/.

37. T. D. Schmittgen, K. J. Livak, Analyzing real-time PCR data by the comparative C-T method. *Nat Protoc* **3**, 1101-1108 (2008).

38. D. Y. Zheng, M. B. Gerstein, The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* **23**, 219-224 (2007).

39. O. H. Tam *et al.*, Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-538 (2008).

40. L. Poliseno *et al.*, A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-U1090 (2010).

41. M. Gerstein, D. Y. Zheng, The real life of pseudogenes. *Sci Am* **295**, 48-55 (2006).

42. T. Kondo *et al.*, Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila Embryogenesis. *Science* **329**, 336-339 (2010).

43. Z. L. Zhang *et al.*, PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-1439 (2006).

44. M. R. Friedlander *et al.*, Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407-415 (2008).

45. M. Kato, A. de Lencastre, Z. Pincus, F. J. Slack, Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol.* **10**, R54 (2009).

46. J. G. Ruby, C. H. Jan, D. P. Bartel, Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-86 (2007).

47. K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, E. C. Lai, The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**, 89-100 (2007).

48. W. Chung *et al.*, Computational and experimental identification of mirtrons in *Drosophila melanogaster* and *Caenorhabditis elegans*. *Genome Res.*, 10.1101/gr.113050.110.

49. P. J. Batista *et al.*, PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C-elegans. *Mol. Cell* **31**, 67-78 (2008).

50. J. M. Claycomb *et al.*, The Argonaute CSR-1 and Its 22G-RNA Cofactors Are Required for Holocentric Chromosome Segregation. *Cell* **139**, 123-134 (2009).

51. C. C. Conine *et al.*, Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3588-3593 (2010).

52. E. de Wit, S. E. V. Linsen, E. Cuppen, E. Berezikov, Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res.* **19**, 2064-2074 (2009).

53. J. I. Gent *et al.*, Distinct Phases of siRNA Synthesis in an Endogenous RNAi Pathway in C. elegans Soma. *Mol. Cell* **37**, 679-689 (2010).

54. J. I. Gent *et al.*, A Caenorhabditis elegans RNA-Directed RNA Polymerase in Sperm Development and Endogenous RNA Interference. *Genetics* **183**, 1297-1314 (2009).

55. J. G. Ruby *et al.*, Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C-elegans. *Cell* **127**, 1193-1207 (2006).

56. M. Stoeckius *et al.*, Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat. Methods* **6**, 745-U716 (2009).

57. J. C. van Wolfswinkel *et al.*, CDE-1 Affects Chromosome Segregation through Uridylation of CSR-1-Bound siRNAs. *Cell* **139**, 135-148 (2009).

58. Full analysis of *C. elegans* mirtrons, http://cbio.mskcc.org/leslielab/mirtrons/ce_mirtrons
59. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154-D158 (2008).
60. Z. J. Lu *et al.*, Prediction and characterization of non-coding RNAs in C. elegans by integrating conservation, secondary structure and high throughput sequencing and array data. *Genome Res.*, 10.1101/gr.110189.110.
61. J. Rozowsky *et al.*, PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66-75 (2009).
62. W. Niu *et al*., Diverse transcription factor binding features revealed by genome-wide ChIP-Seq in C. elegans. *Genome Res.*, 10.1101/gr.114587.110.
63. P. V. Kharchenko, M. Y. Tolstorukov, P. J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351-1359 (2008).
64. M. Zhong *et al.*, Genome-Wide Identification of Binding Sites Defines Distinct Functions for Caenorhabditis elegans PHA-4/FOXA in Development and Environmental Response. *PLoS Genet.* **6**, e1000848 (2010).
65. G. D. Stormo, DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
66. Aggregation and Correlation Toolbox, http://act.gersteinlab.org
67. S. M. Johnson, S. Y. Lin, F. J. Slack, The time of appearance of the C-elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev. Biol.* **259**, 364-379 (2003).
68. S. Ercan *et al.*, X chromosome repression by localization of the C-elegans dosage compensation machinery to sites of transcription initiation. *Nature Genet.* **39**, 403-408 (2007).
69. P. J. Roy, J. M. Stuart, J. Lund, S. K. Kim, Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418**, 975-979 (2002).
70. C. A. Grove *et al.*, A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. *Cell* **138**, 314-327 (2009).
71. F. Pauli, Y. Y. Liu, Y. A. Kim, P. J. Chen, S. K. Kim, Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C-elegans. *Development* **133**, 287-295 (2006).
72. J. S. Gilleard, Y. Shafi, J. D. Barry, J. D. McGhee, ELT-3: A Caenorhabditis elegans GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev. Biol.* **208**, 265-280 (1999).
73. J. Gaudet, S. E. Mango, Regulation of organogenesis by the Caenorhabditis elegans, FoxA protein PHA-41. *Science* **295**, 821-825 (2002).
74. R. S. Kamath *et al.*, Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-237 (2003).
75. Cytoscape, http://cytoscape.org
76. The Yale Network Analyzer (tYNA), http://tyna.gersteinlab.org
77. N. Simonis *et al.*, Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods* **6**, 47-54 (2009).
78. A. Krek *et al.*, Combinatorial microRNA target predictions. *Nature Genet.* **37**, 495-500 (2005).
79. S. Lall *et al.*, A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.* **16**, 460-471 (2006).
80. AceView, http://www.aceview.org
81. B. P. Lewis, C. B. Burge, D. P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20 (2005).
82. A. Stark, J. Brennecke, N. Bushati, R. B. Russell, S. M. Cohen, Animal microRNAs confer robustness to gene expression and have a significant impact on 3 ' UTR evolution. *Cell* **123**, 1133-1146 (2005).
83. N. Rajewsky, microRNA target predictions in animals. *Nature Genet.* **38**, S8-S13 (2006).
84. M. Selbach *et al.*, Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58-63 (2008).

85. M. Hall *et al.*, The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11**, 10-18 (2009).

86. Methods used to build six-way nematode genome alignments, http://genome.ucsc.edu/cgi-bin/hgc?g=multiz6way

87. Genome sequence of *Caenorhabditis briggsae*, http://genome.wustl.edu/genomes/view/caenorhabditis_briggsae

88. Genome sequence of *Caenorhabditis remanei*, http://genome.wustl.edu/genomes/view/caenorhabditis_remanei

89. Genome sequence of *Caenorhabditis brenneri*, http://genome.wustl.edu/genomes/view/caenorhabditis_brenneri

90. Genome sequence of *Pristionchus pacificus*, http://genome.wustl.edu/genomes/view/pristionchus_pacificus_var._california

91. Genome sequence of *Caenorhabditis japonica*, http://genome.wustl.edu/genomes/view/caenorhabditis_japonica

92. Phylogenetic analysis with space/time models (PHAST) software, http://compgen.bscb.cornell.edu/phast

93. K. Kiontke, D. H. A. Fitch, The Phylogenetic relationships of *Caenorhabditis* and other rhabditids (August 11, 2005), *WormBook,* ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.11.1.

94. The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

95. P. J. Bickel, N. Boley, J. B. Brown, H. Huang, N. Zhang, Subsampling methods for genomic inference. *Annals Applied Stat.* **0**, 1-41 (2010).

96. K. D. Pruitt *et al.*, The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316-1323 (2009).

97. T. W. Harris *et al.*, WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* **38**, D463-D467 (2010).

98. A. Barski *et al.*, High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

99. Y. Zhang, H. Shin, J. S. Song, Y. Lei, X. S. Liu, Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq. *BMC Genomics* **9**, - (2008).

100. S. Meader, C. P. Ponting, G. Lunter, Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* **20**, 1335-1343 (2010).

101. C. elegans Sequencing Consortium, Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).

102. P. Vaglio *et al.*, WorfDB: the Caenorhabditis elegans ORFeome database. *Nucleic Acids Res.* **31**, 237-240 (2003).

103. Integrative Genomics Viewer, http://www.broadinstitute.org/igv/v1.2

104. SHRiMP aligner v1.3, http://compbio.cs.toronto.edu/shrimp

105. D. L. Wheeler *et al.*, Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13-D21 (2008).

106. T. Fukushige, M. Krause, The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early C-elegans embryos. *Development* **132**, 1795-1805 (2005).

107. X. Liu *et al.*, Analysis of Cell Fate from Single-Cell Gene Expression Profiles in C. elegans. *Cell* **139**, 623-633 (2009).

108. E. Portales-Casamar *et al.*, JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105-D110 (2010).

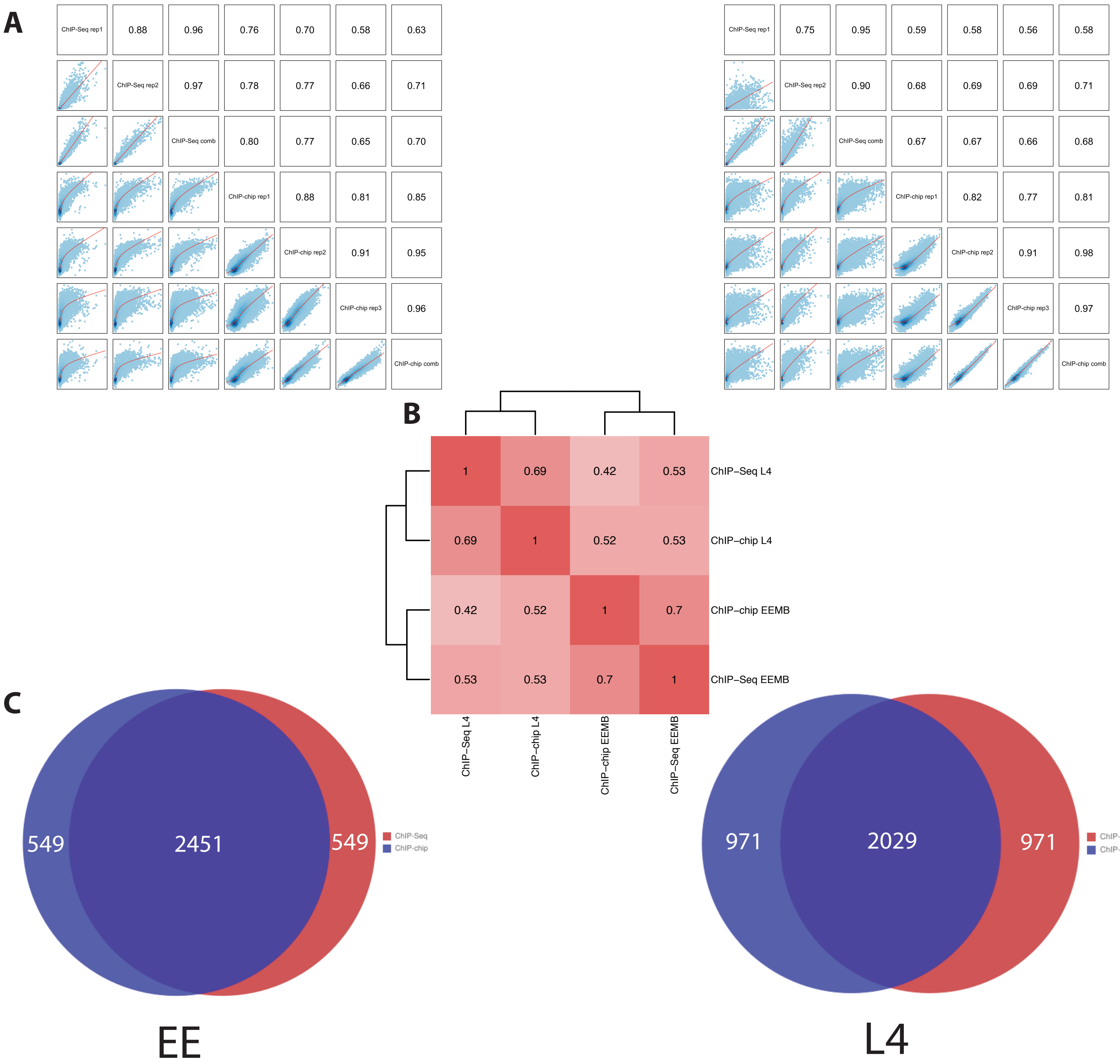
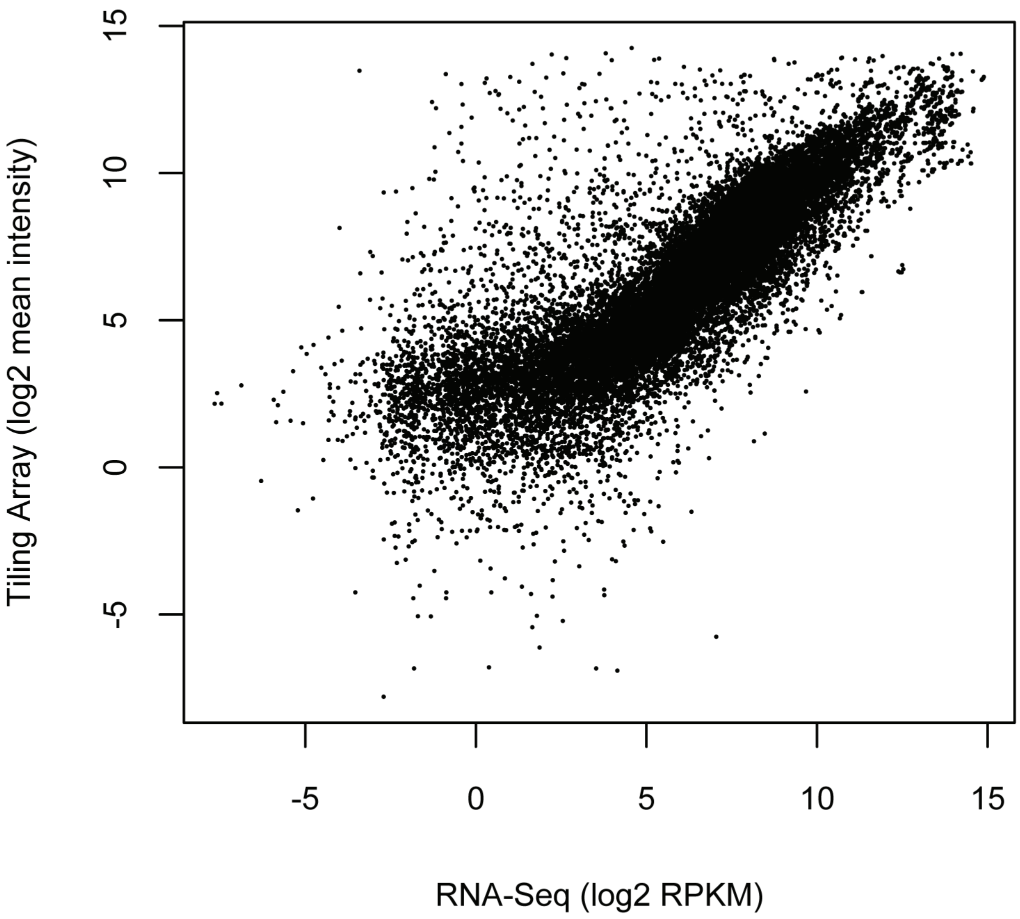

A
ChIP-Seq rep1
0.88
0.96
0.76
0.70
0.58
0.63
ChIP-Seq rep2
0.97
0.78
0.77
0.66
0.71
ChIP-Seq comb
0.80
0.77
0.65
0.70
ChIP-chip rep1
0.88
0.81
0.85
ChIP-chip rep2
0.91
0.95
ChIP-chip rep3
0.96
ChIP-chip comb
ChIP-Seq rep1
0.75
0.95
0.59
0.58
0.56
0.58
ChIP-Seq rep2
0.90
0.68
0.69
0.69
0.71
ChIP-Seq comb
0.67
0.67
0.66
0.68
ChIP-chip rep1
0.82
0.77
0.81
ChIP-chip rep2
0.91
0.98
ChIP-chip rep3
0.97
ChIP-chip comb
B
1
0.69
0.42
0.53
ChIP–Seq L4
0.69
1
0.52
0.53
ChIP–chip L4
0.42
0.52
1
0.7
ChIP–chip EEMB
0.53
0.53
0.7
1
ChIP–Seq EEMB
ChIP–Seq L4
ChIP–chip L4
ChIP–chip EEMB
ChIP–Seq EEMB
C
549
2451
549
ChIP-Seq
ChIP-chip
EE
971
2029
971
ChIP-Seq
ChIP-chip
L4

Fig. S1



15
10
5
0
-5
Tiling Array (log2 mean intensity)
-5
0
5
10
15
RNA-Seq (log2 RPKM)

Fig. S2

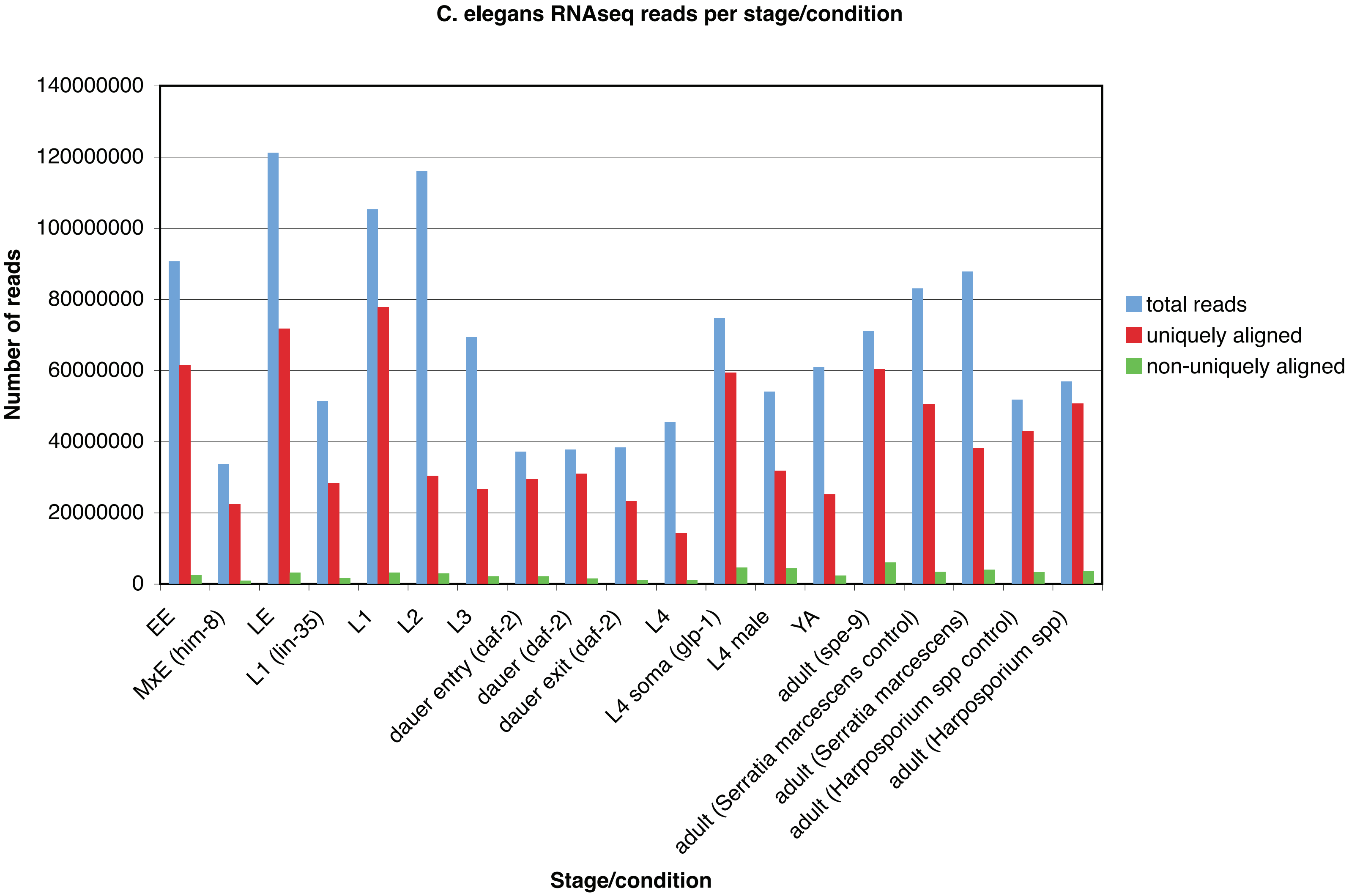

C. elegans RNAseq reads per stage/condition
Number of reads
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
EE
MxE (him-8)
LE
L1 (lin-35)
L1
L2
L3
dauer entry (daf-2)
dauer (daf-2)
dauer exit (daf-2)
L4
L4 soma (glp-1)
L4 male
YA
adult (spe-9)
adult (Serratia marcescens control)
adult (Serratia marcescens)
adult (Harposporium spp control)
adult (Harposporium spp)
Stage/condition
🟦 total reads
🟥 uniquely aligned
🟩 non-uniquely aligned

Fig. S3

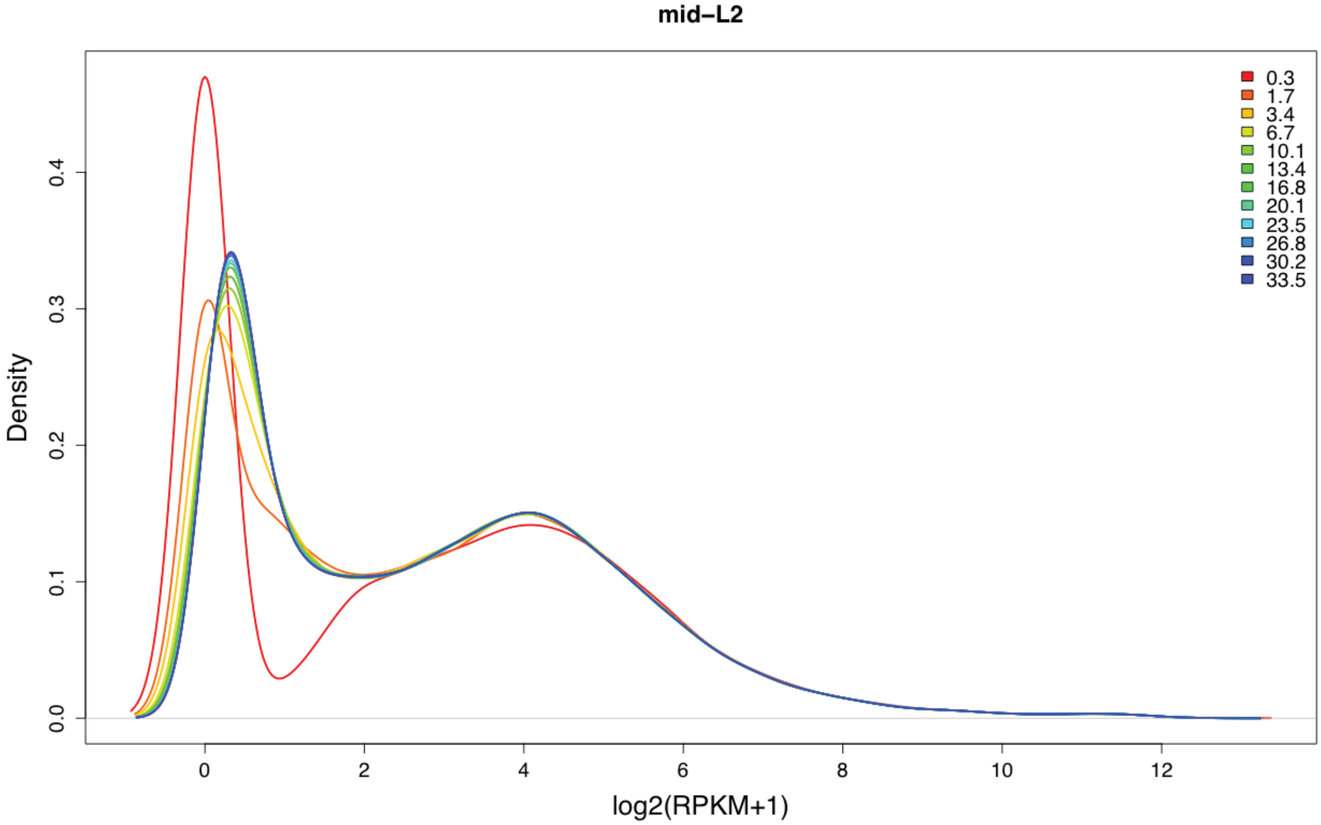

mid–L2
0.3
1.7
3.4
6.7
10.1
13.4
16.8
20.1
23.5
26.8
30.2
33.5
Density
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
12
log2(RPKM+1)

Fig. S4A

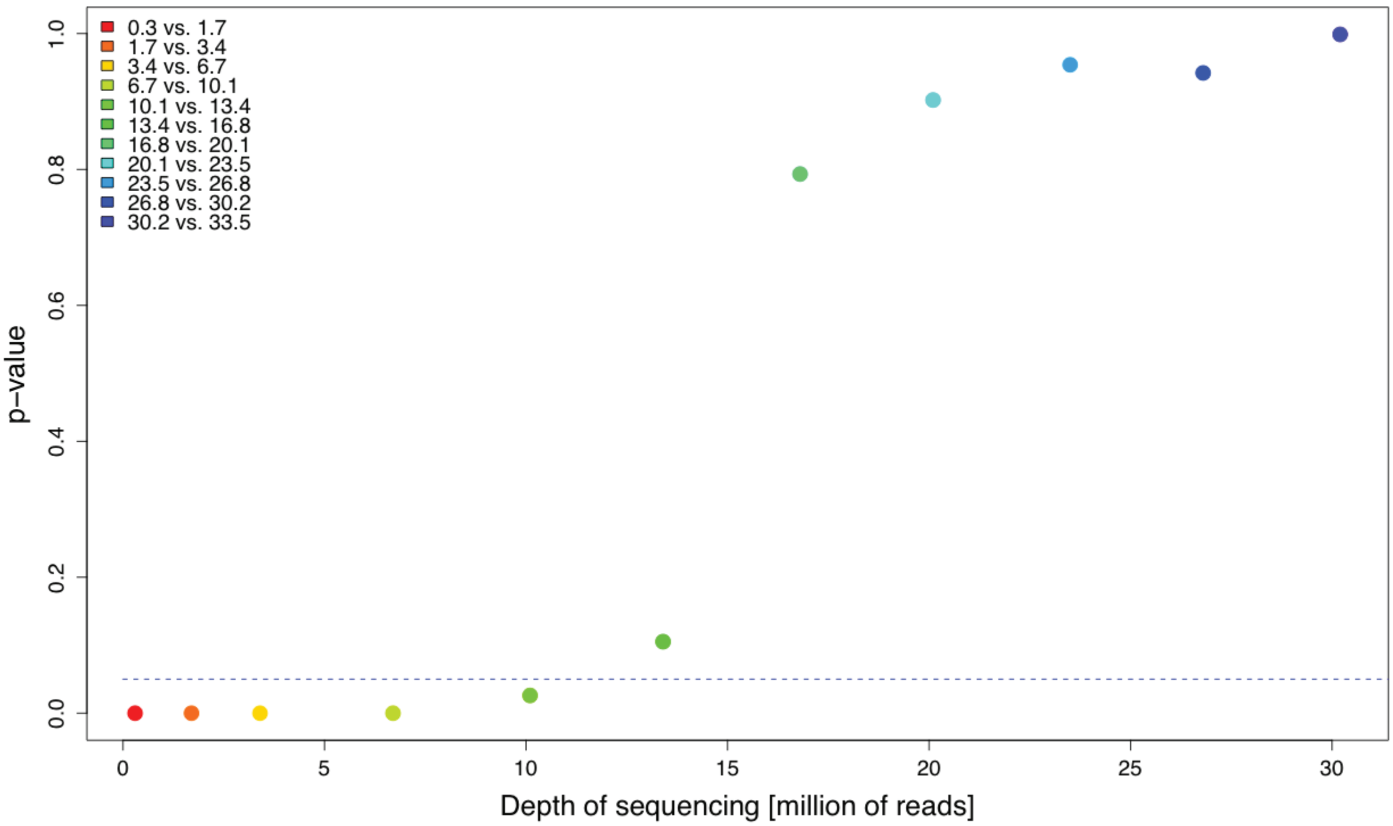

0.3 vs. 1.7
1.7 vs. 3.4
3.4 vs. 6.7
6.7 vs. 10.1
10.1 vs. 13.4
13.4 vs. 16.8
16.8 vs. 20.1
20.1 vs. 23.5
23.5 vs. 26.8
26.8 vs. 30.2
30.2 vs. 33.5
p−value
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Depth of sequencing [million of reads]

Fig. S4B

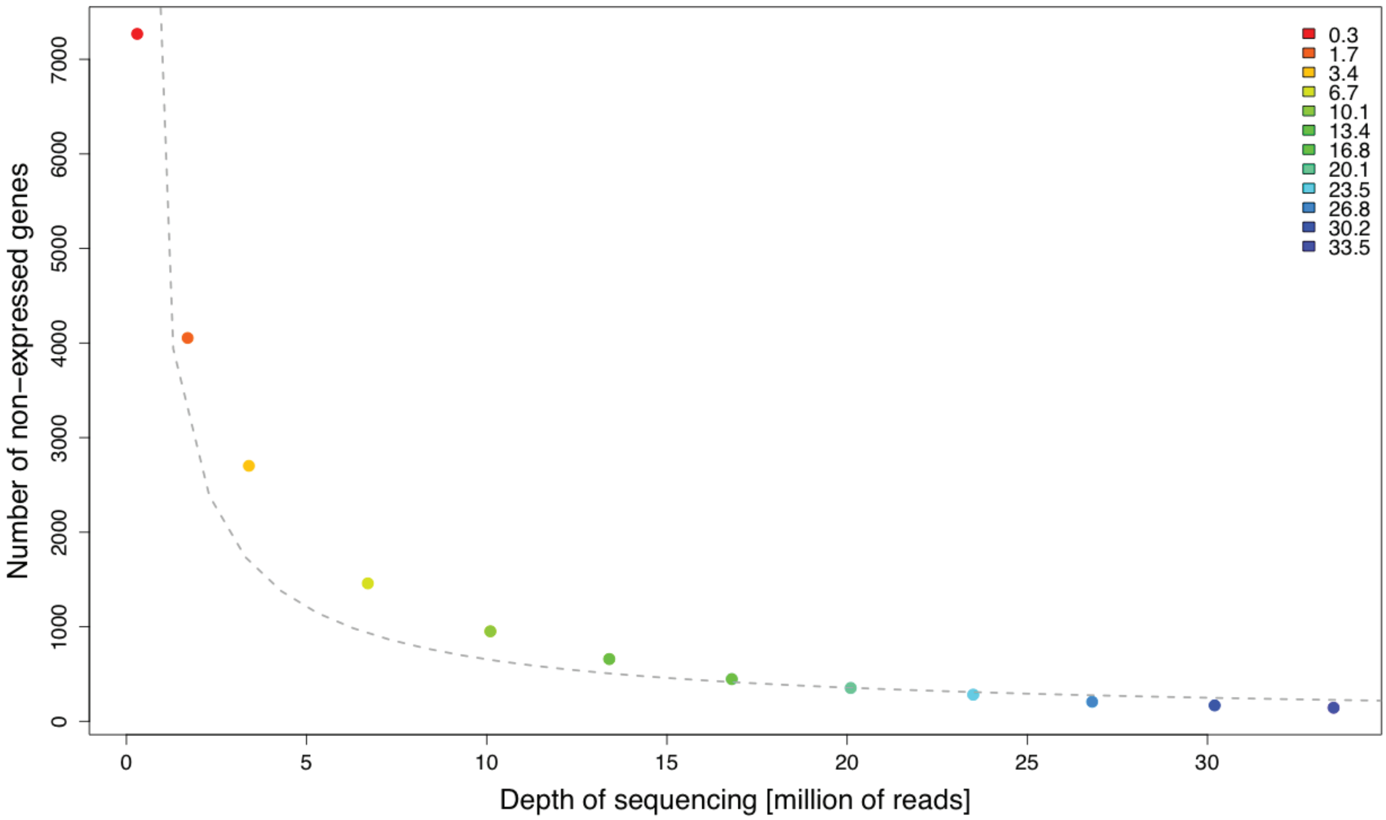

Number of non−expressed genes
0
1000
2000
3000
4000
5000
6000
7000
0
5
10
15
20
25
30
Depth of sequencing [million of reads]
0.3
1.7
3.4
6.7
10.1
13.4
16.8
20.1
23.5
26.8
30.2
33.5

Fig. S4C

Transcript building

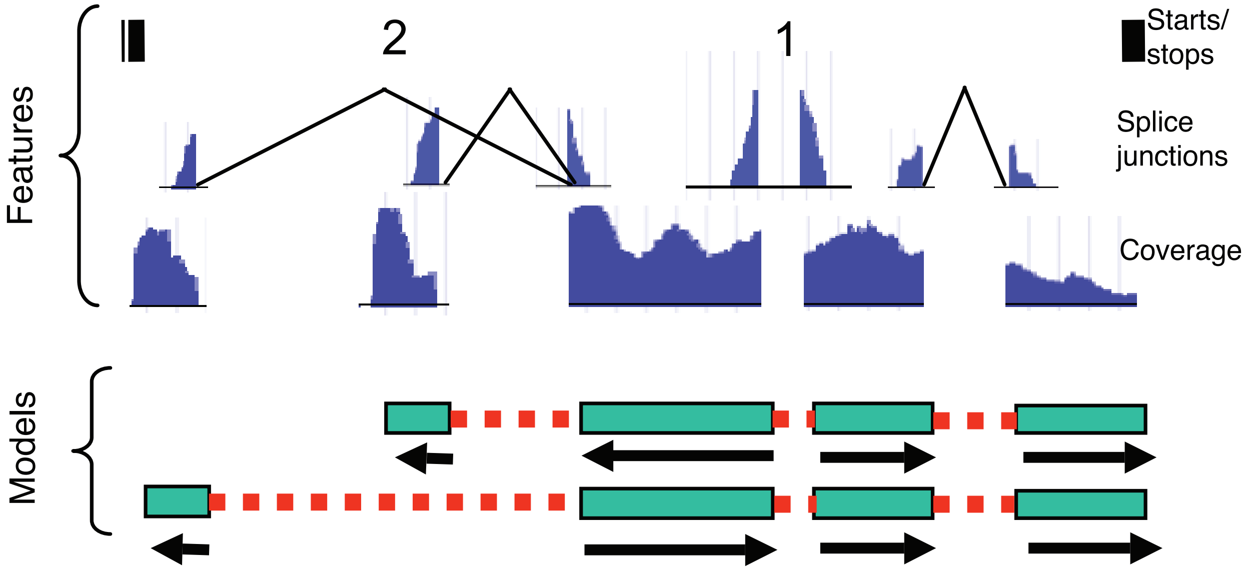


Fig. S5

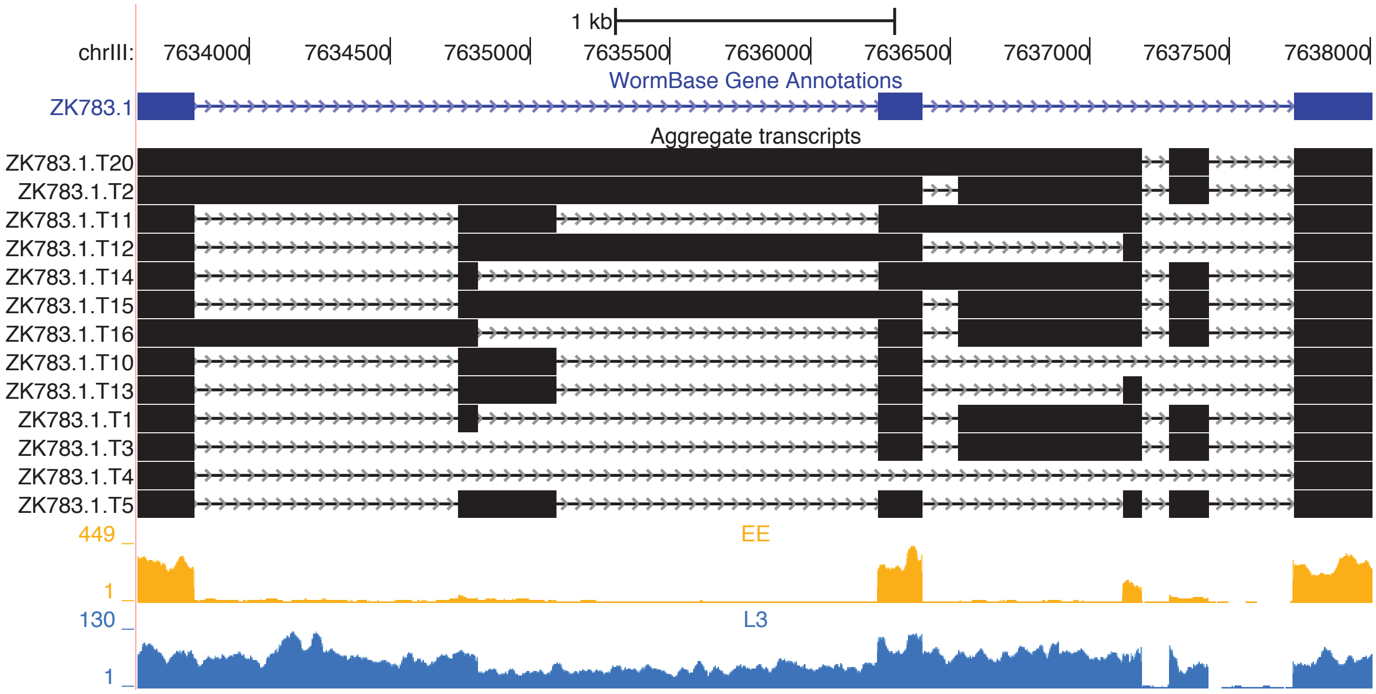

1 kb
chrIII:
7634000
7634500
7635000
7635500
7636000
7636500
7637000
7637500
7638000
WormBase Gene Annotations
ZK783.1
Aggregate transcripts
ZK783.1.T20
ZK783.1.T2
ZK783.1.T11
ZK783.1.T12
ZK783.1.T14
ZK783.1.T15
ZK783.1.T16
ZK783.1.T10
ZK783.1.T13
ZK783.1.T1
ZK783.1.T3
ZK783.1.T4
ZK783.1.T5
449 _
EE
1 _
130 _
L3
1 _

Fig. S6

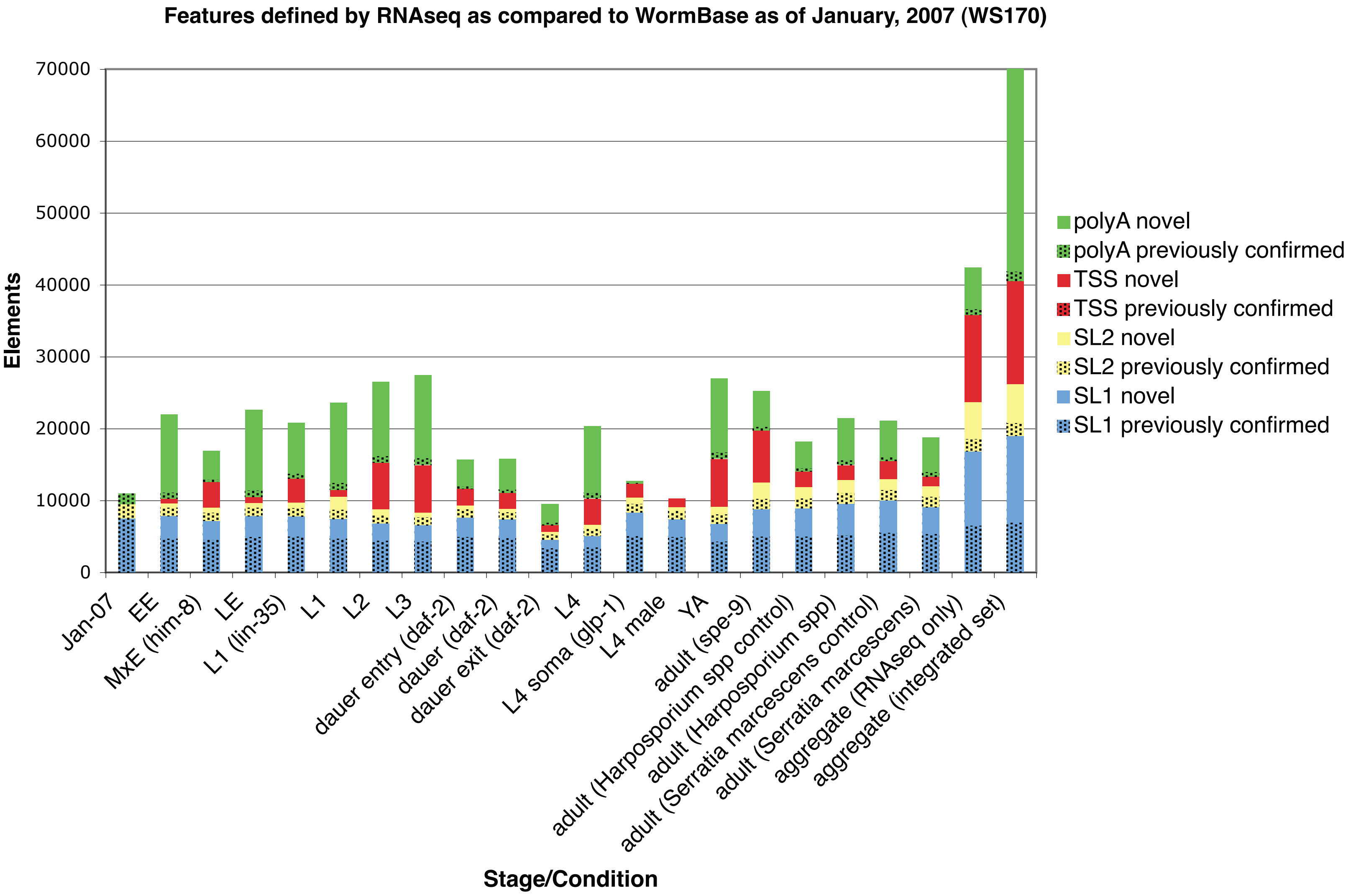

Features defined by RNAseq as compared to WormBase as of January, 2007 (WS170)
Elements
70000
60000
50000
40000
30000
20000
10000
0
Jan-07
EE
MxE (him-8)
LE
L1 (lin-35)
L1
L2
L3
dauer entry (daf-2)
dauer (daf-2)
dauer exit (daf-2)
L4
L4 soma (glp-1)
L4 male
YA
adult (spe-9)
adult (Harposporium spp control)
adult (Harposporium spp)
adult (Serratia marcescens control)
adult (Serratia marcescens)
aggregate (RNAseq only)
aggregate (integrated set)
Stage/Condition
polyA novel
polyA previously confirmed
TSS novel
TSS previously confirmed
SL2 novel
SL2 previously confirmed
SL1 novel
SL1 previously confirmed

Fig. S7

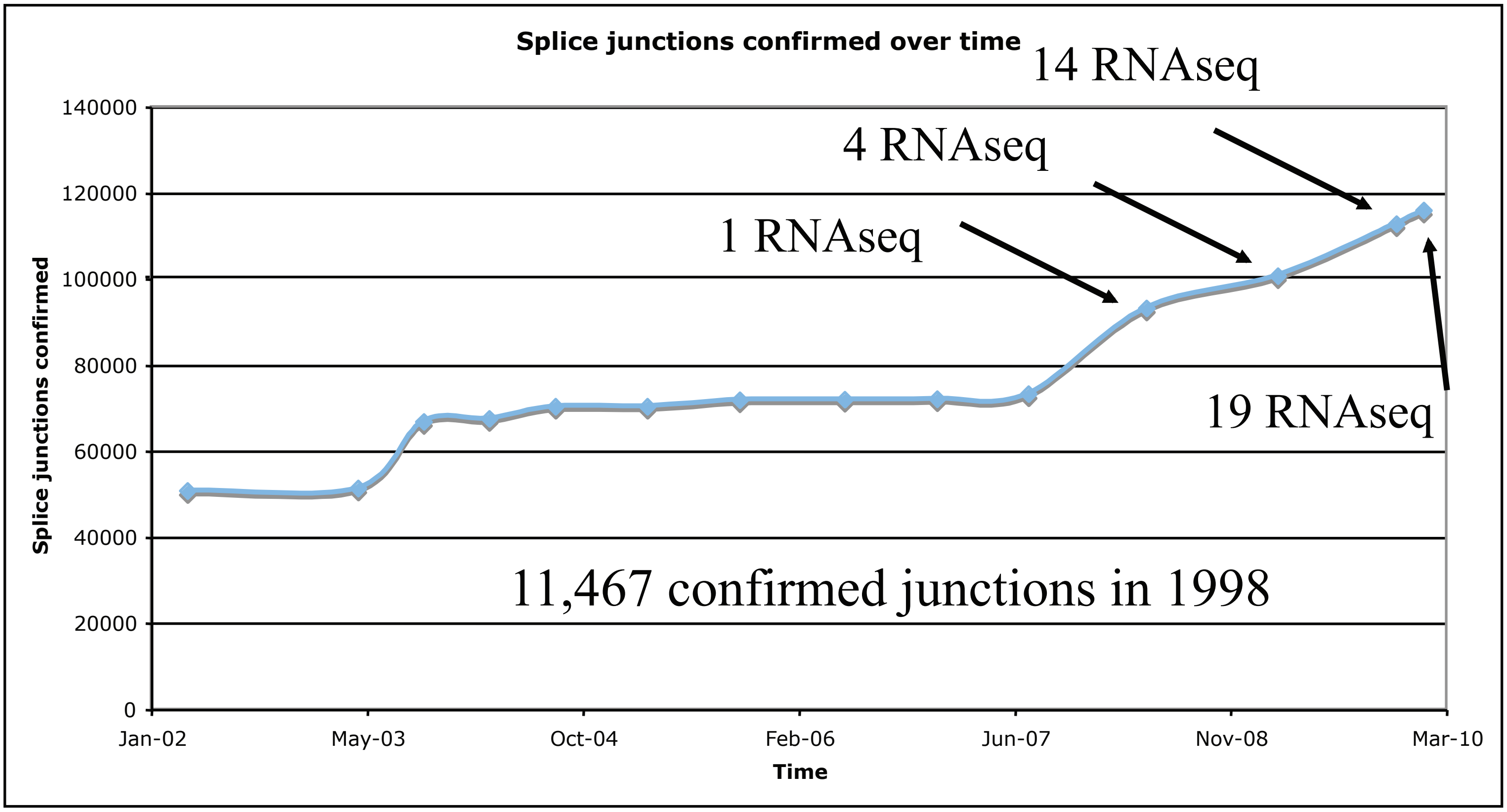

Splice junctions confirmed over time
14 RNAseq
4 RNAseq
1 RNAseq
19 RNAseq
11,467 confirmed junctions in 1998
Splice junctions confirmed
140000
120000
100000
80000
60000
40000
20000
0
Jan-02
May-03
Oct-04
Feb-06
Jun-07
Nov-08
Mar-10
Time

Fig. S8

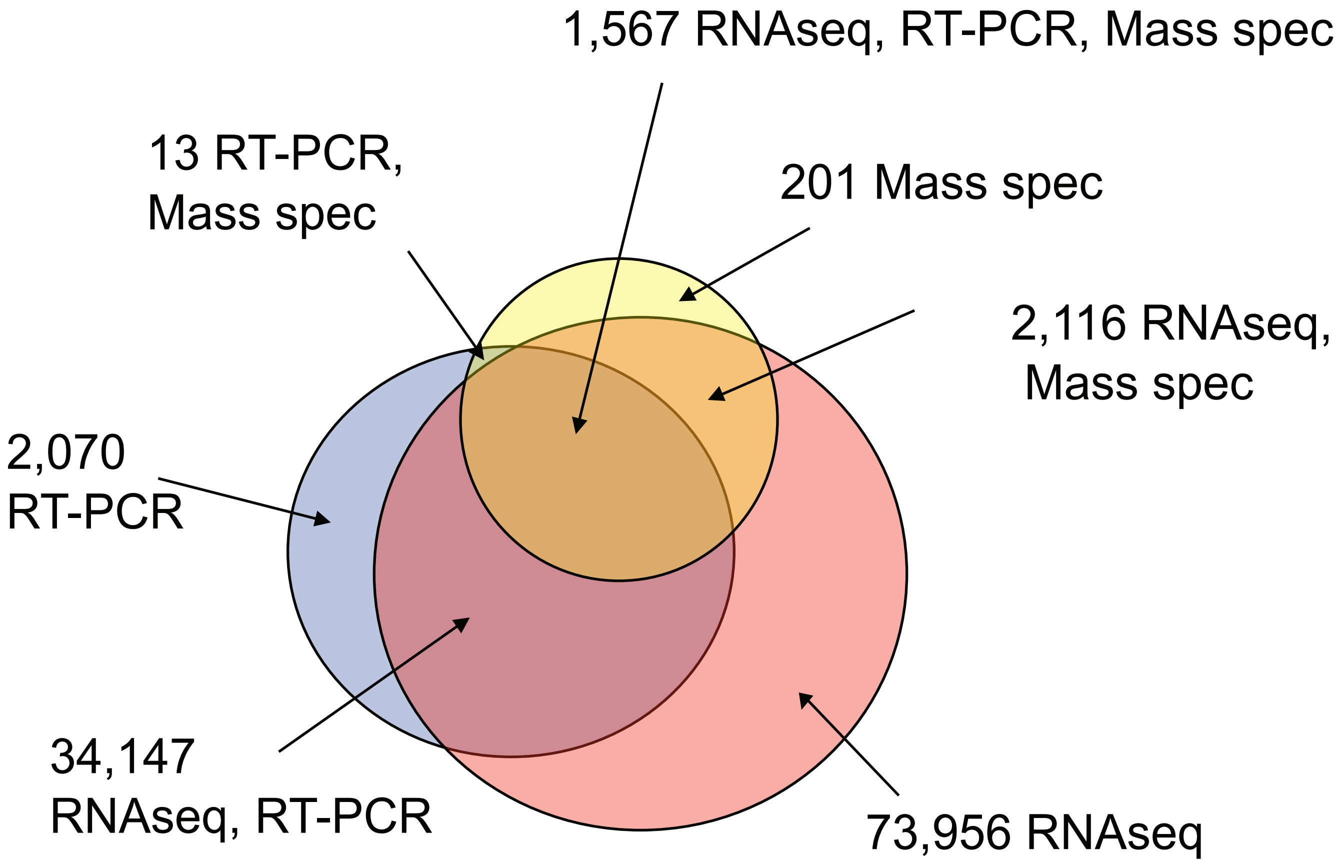

1,567 RNAseq, RT-PCR, Mass spec
13 RT-PCR,
Mass spec
201 Mass spec
2,116 RNAseq,
Mass spec
2,070
RT-PCR
34,147
RNAseq, RT-PCR
73,956 RNAseq

Fig. S9

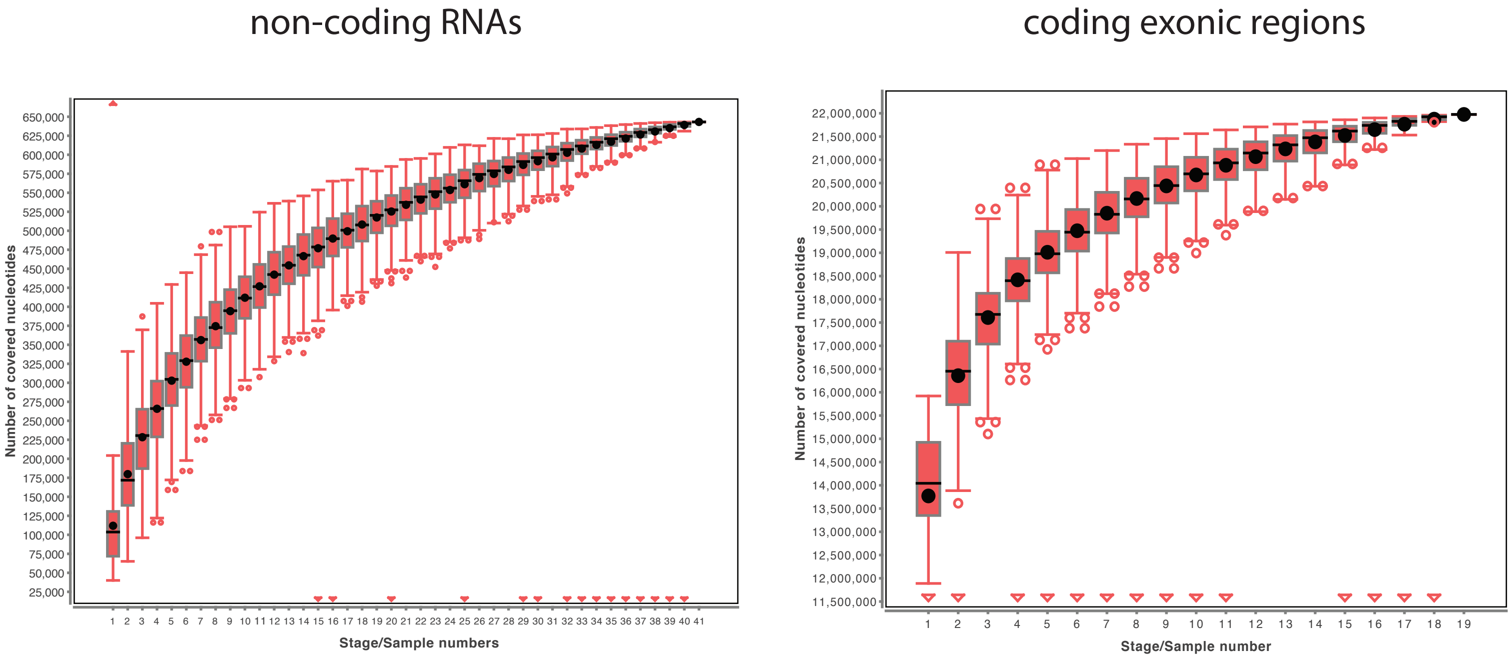

non-coding RNAs
Number of covered nucleotides
650,000
625,000
600,000
575,000
550,000
525,000
500,000
475,000
450,000
425,000
400,000
375,000
350,000
325,000
300,000
275,000
250,000
225,000
200,000
175,000
150,000
125,000
100,000
75,000
50,000
25,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
Stage/Sample numbers
coding exonic regions
Number of covered nucleotides
22,000,000
21,500,000
21,000,000
20,500,000
20,000,000
19,500,000
19,000,000
18,500,000
18,000,000
17,500,000
17,000,000
16,500,000
16,000,000
15,500,000
15,000,000
14,500,000
14,000,000
13,500,000
13,000,000
12,500,000
12,000,000
11,500,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Stage/Sample number

Fig. S10

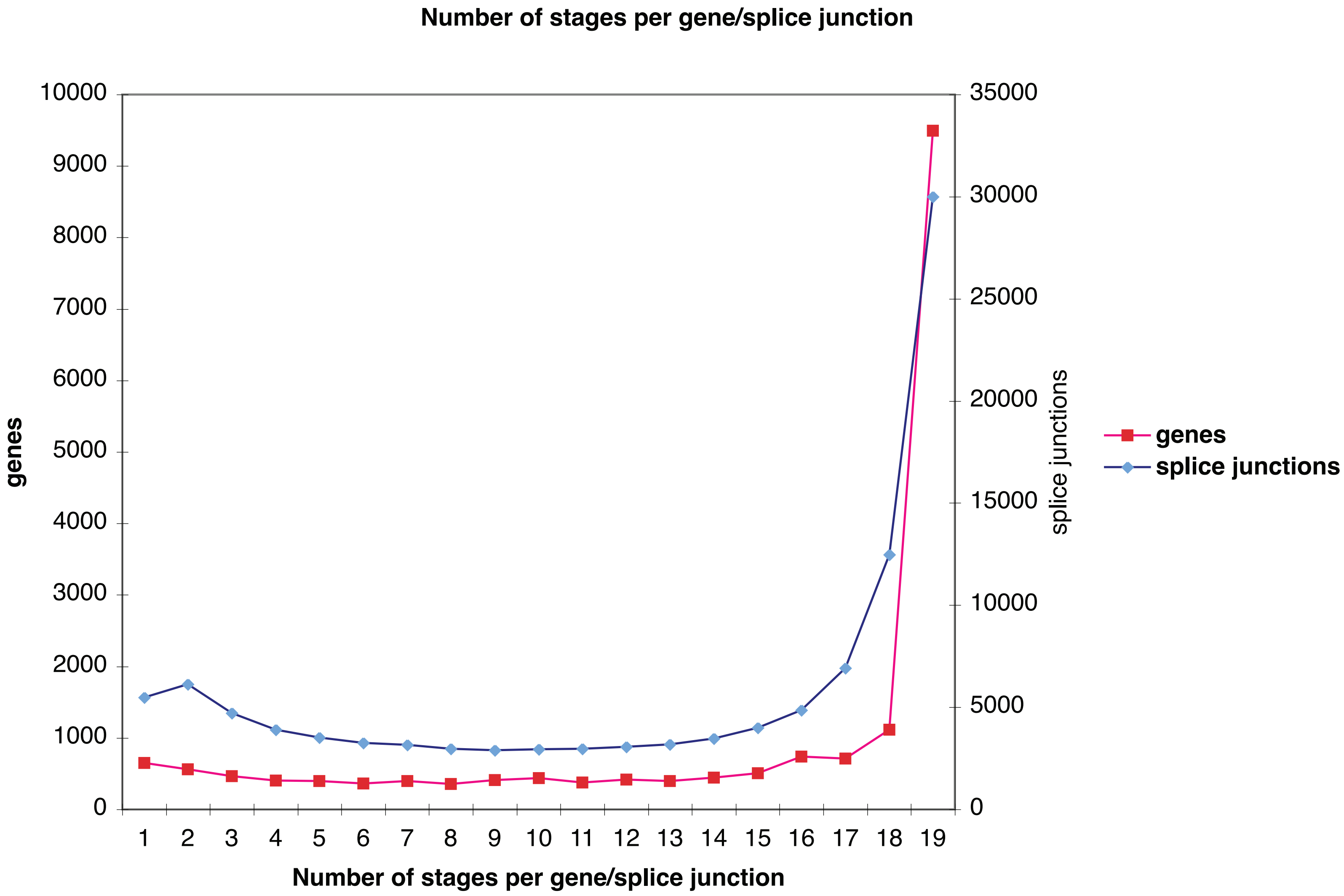

Number of stages per gene/splice junction
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
genes
35000
30000
25000
20000
15000
10000
5000
0
splice junctions
genes
splice junctions
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Number of stages per gene/splice junction

Fig. S11

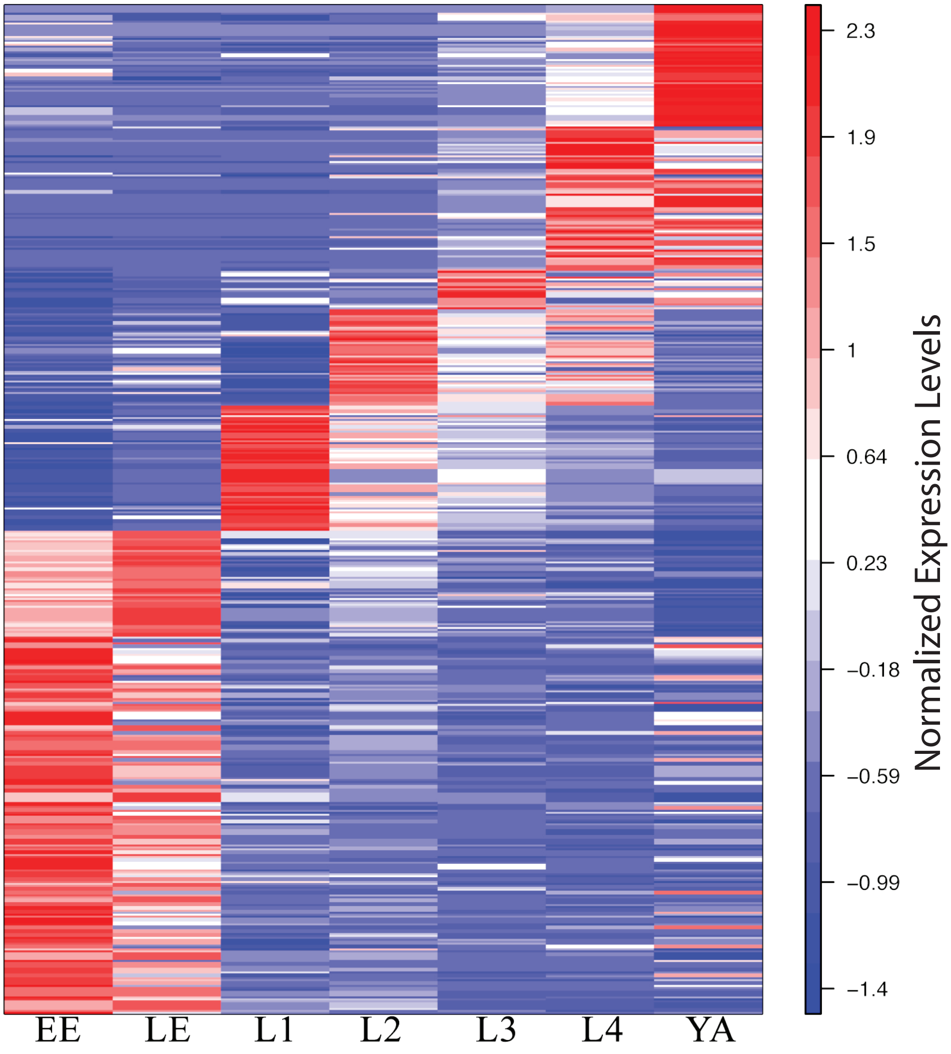
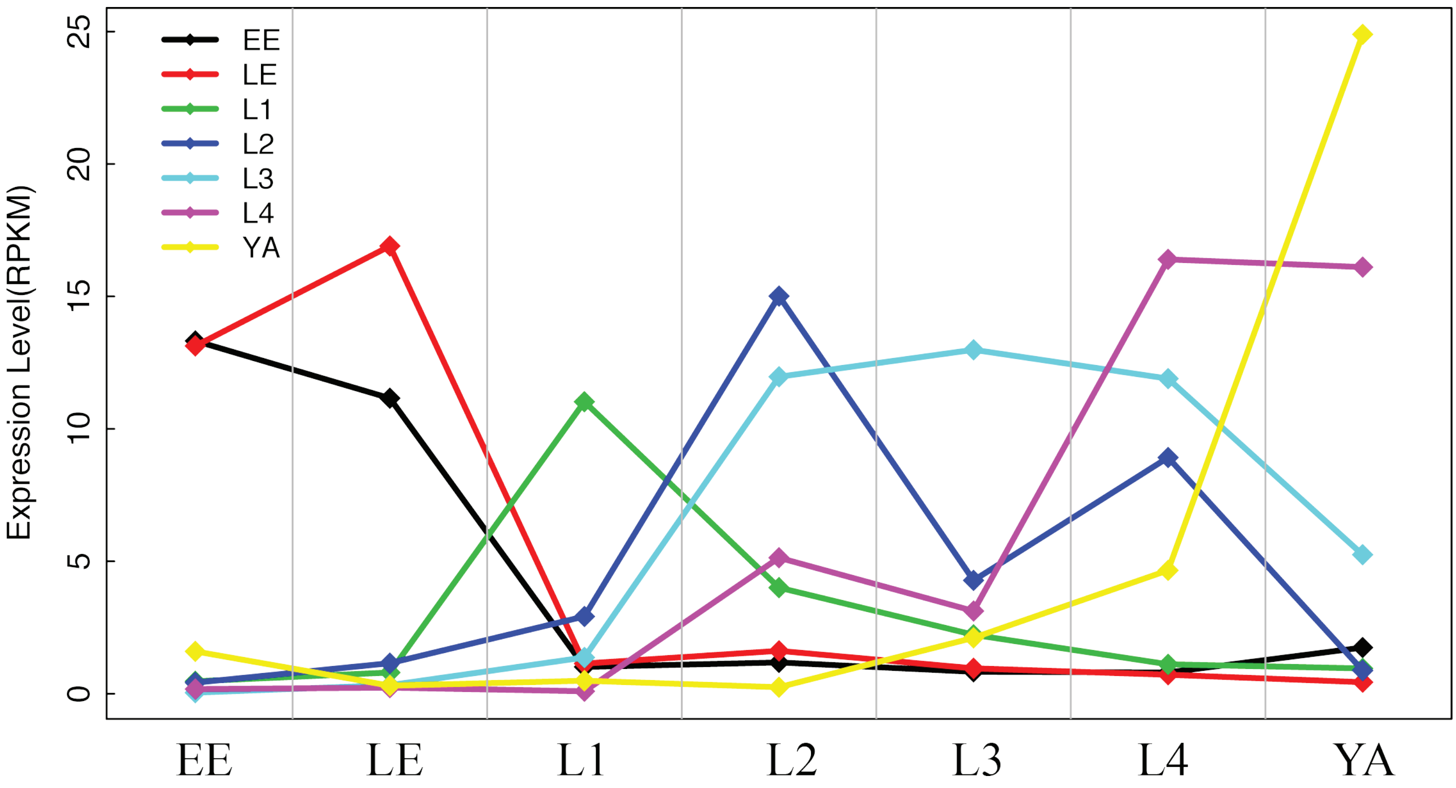

EE
LE
L1
L2
L3
L4
YA
2.3
1.9
1.5
1
0.64
0.23
−0.18
−0.59
−0.99
−1.4
Normalized Expression Levels

Fig. S12A



Expression Level(RPKM)
0
5
10
15
20
25
EE
LE
L1
L2
L3
L4
YA
EE
LE
L1
L2
L3
L4
YA

Fig. S12B

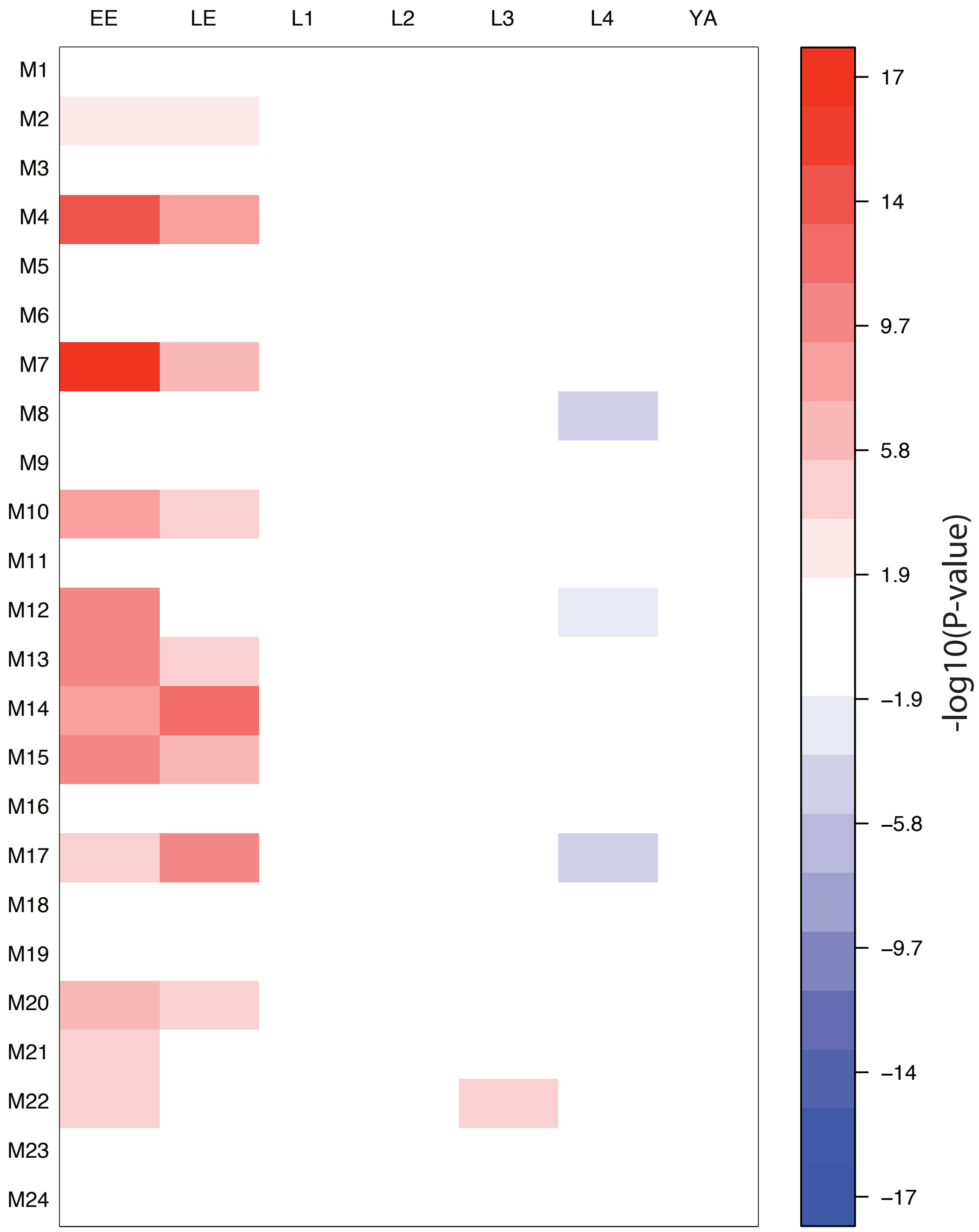
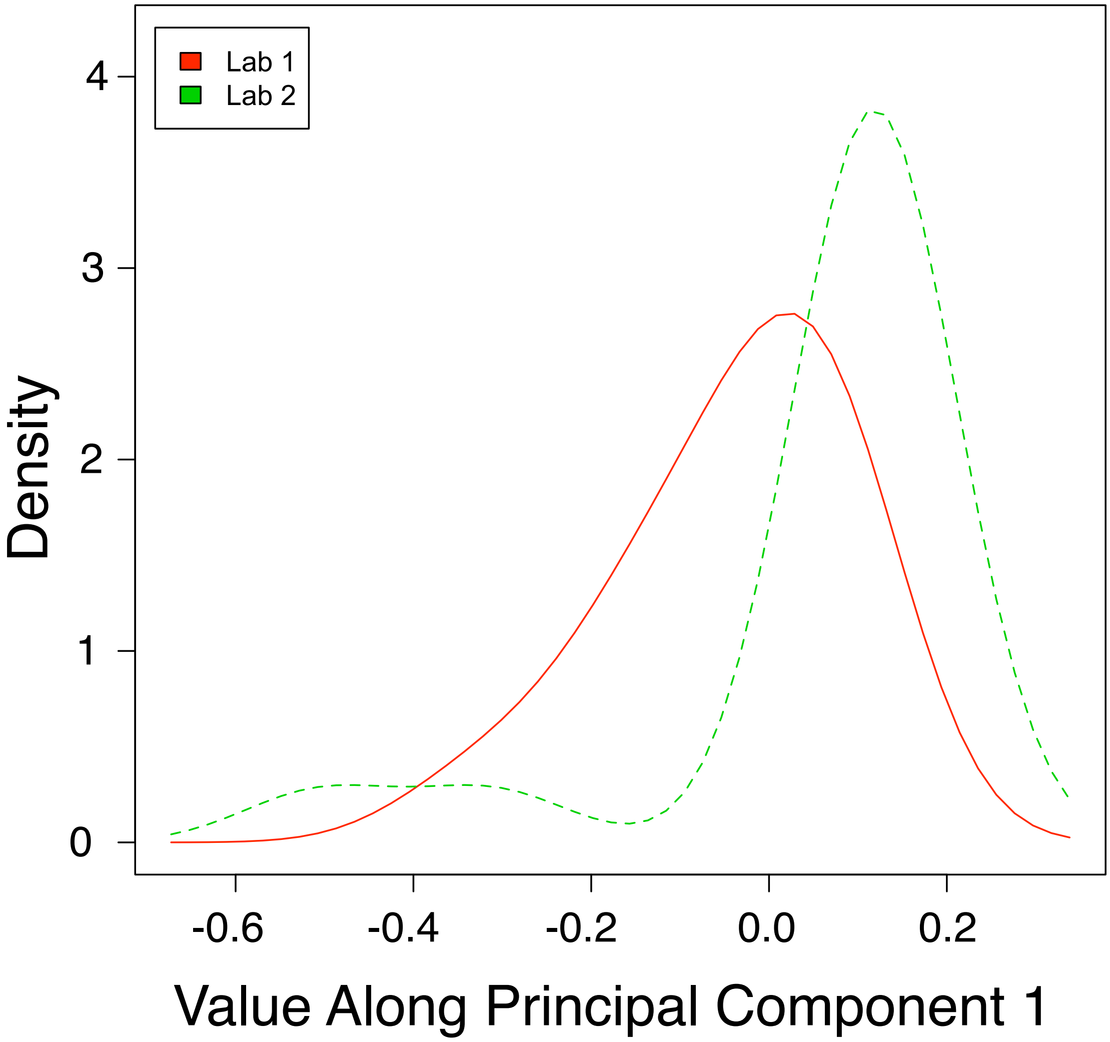

EE
LE
L1
L2
L3
L4
YA
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
M21
M22
M23
M24
17
14
9.7
5.8
1.9
−1.9
−5.8
−9.7
−14
−17
-log10(P-value)

Fig. S12C



Lab 1
Lab 2
4
3
2
1
0
Density
-0.6
-0.4
-0.2
0.0
0.2
Value Along Principal Component 1

Fig. S13

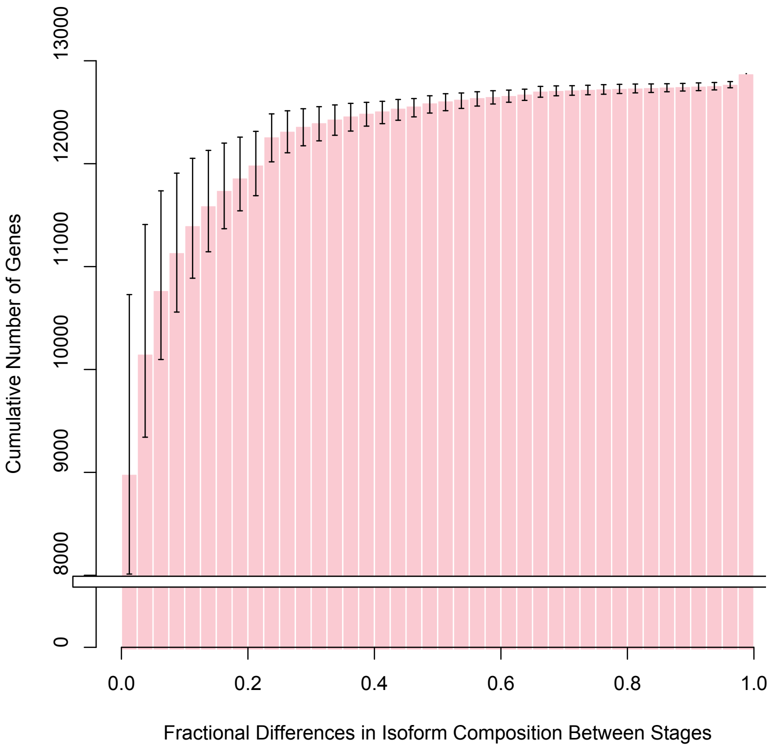

Cumulative Number of Genes
13000
12000
11000
10000
9000
8000
0
0.0
0.2
0.4
0.6
0.8
1.0
Fractional Differences in Isoform Composition Between Stages

Fig. S14

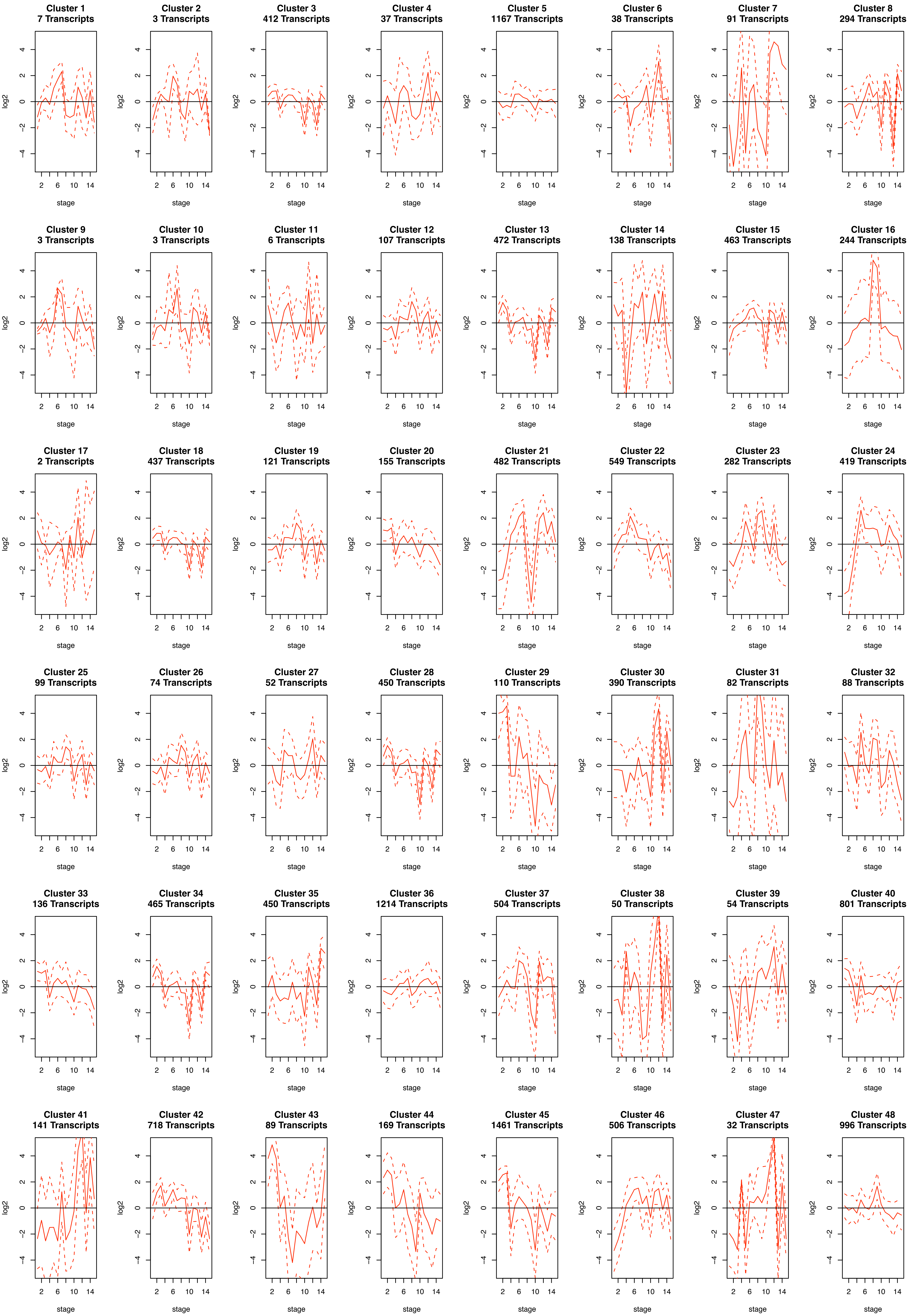

Cluster 1
7 Transcripts
Cluster 2
3 Transcripts
Cluster 3
412 Transcripts
Cluster 4
37 Transcripts
Cluster 5
1167 Transcripts
Cluster 6
38 Transcripts
Cluster 7
91 Transcripts
Cluster 8
294 Transcripts
Cluster 9
3 Transcripts
Cluster 10
3 Transcripts
Cluster 11
6 Transcripts
Cluster 12
107 Transcripts
Cluster 13
472 Transcripts
Cluster 14
138 Transcripts
Cluster 15
463 Transcripts
Cluster 16
244 Transcripts
Cluster 17
2 Transcripts
Cluster 18
437 Transcripts
Cluster 19
121 Transcripts
Cluster 20
155 Transcripts
Cluster 21
482 Transcripts
Cluster 22
549 Transcripts
Cluster 23
282 Transcripts
Cluster 24
419 Transcripts
Cluster 25
99 Transcripts
Cluster 26
74 Transcripts
Cluster 27
52 Transcripts
Cluster 28
450 Transcripts
Cluster 29
110 Transcripts
Cluster 30
390 Transcripts
Cluster 31
82 Transcripts
Cluster 32
88 Transcripts
Cluster 33
136 Transcripts
Cluster 34
465 Transcripts
Cluster 35
450 Transcripts
Cluster 36
1214 Transcripts
Cluster 37
504 Transcripts
Cluster 38
50 Transcripts
Cluster 39
54 Transcripts
Cluster 40
801 Transcripts
Cluster 41
141 Transcripts
Cluster 42
718 Transcripts
Cluster 43
89 Transcripts
Cluster 44
169 Transcripts
Cluster 45
1461 Transcripts
Cluster 46
506 Transcripts
Cluster 47
32 Transcripts
Cluster 48
996 Transcripts
log2
-4 -2 0 2 4
stage
2 6 10 14

Fig. S15

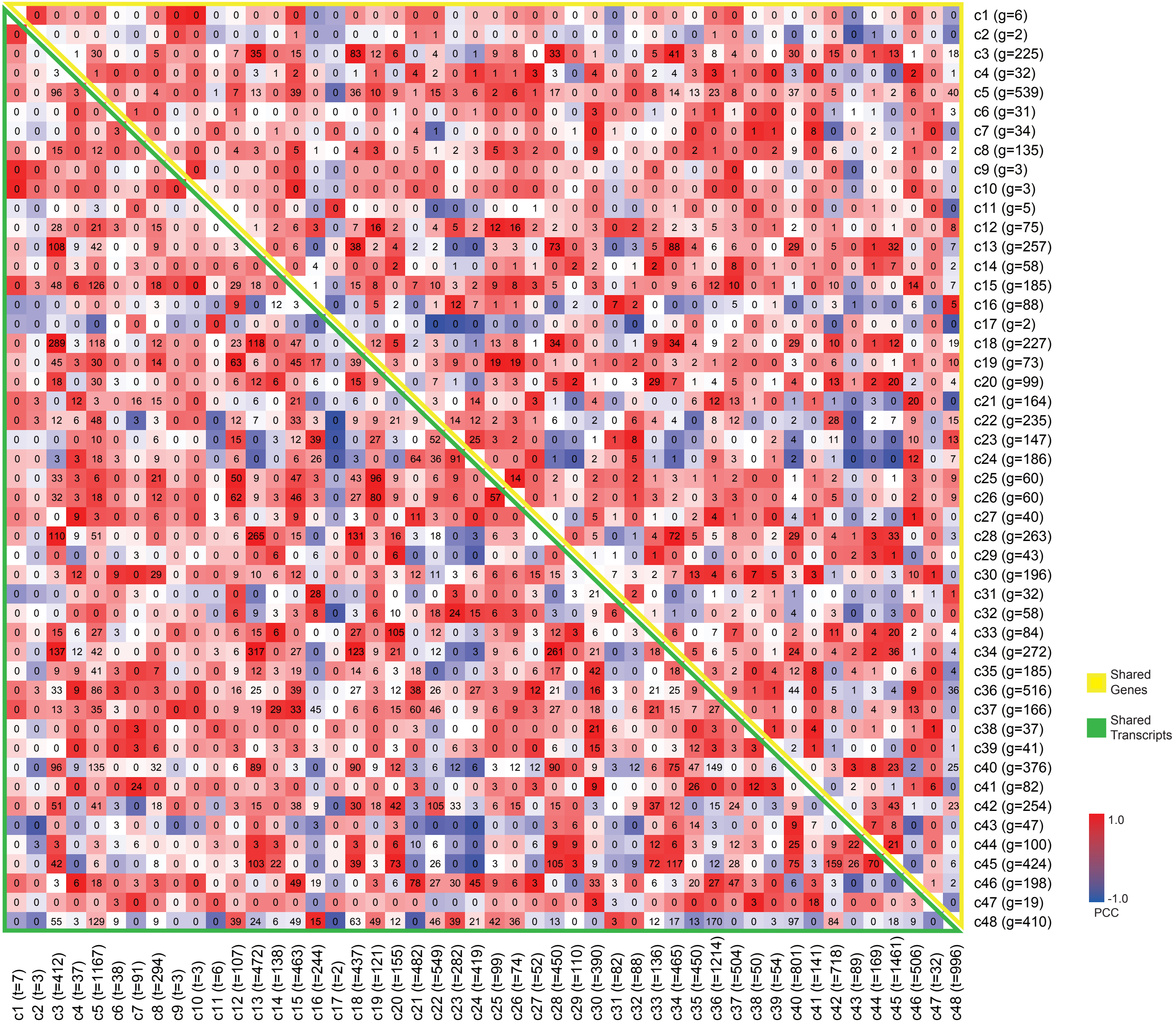

c1 (g=6)
c2 (g=2)
c3 (g=225)
c4 (g=32)
c5 (g=539)
c6 (g=31)
c7 (g=34)
c8 (g=135)
c9 (g=3)
c10 (g=3)
c11 (g=5)
c12 (g=75)
c13 (g=257)
c14 (g=58)
c15 (g=185)
c16 (g=88)
c17 (g=2)
c18 (g=227)
c19 (g=73)
c20 (g=99)
c21 (g=164)
c22 (g=235)
c23 (g=147)
c24 (g=186)
c25 (g=60)
c26 (g=60)
c27 (g=40)
c28 (g=263)
c29 (g=43)
c30 (g=196)
c31 (g=32)
c32 (g=58)
c33 (g=84)
c34 (g=272)
c35 (g=185)
c36 (g=516)
c37 (g=166)
c38 (g=37)
c39 (g=41)
c40 (g=376)
c41 (g=82)
c42 (g=254)
c43 (g=47)
c44 (g=100)
c45 (g=424)
c46 (g=198)
c47 (g=19)
c48 (g=410)
c1 (t=7)
c2 (t=3)
c3 (t=412)
c4 (t=37)
c5 (t=1167)
c6 (t=38)
c7 (t=91)
c8 (t=294)
c9 (t=3)
c10 (t=3)
c11 (t=6)
c12 (t=107)
c13 (t=472)
c14 (t=138)
c15 (t=463)
c16 (t=244)
c17 (t=2)
c18 (t=437)
c19 (t=121)
c20 (t=155)
c21 (t=482)
c22 (t=549)
c23 (t=282)
c24 (t=419)
c25 (t=99)
c26 (t=74)
c27 (t=52)
c28 (t=450)
c29 (t=110)
c30 (t=390)
c31 (t=82)
c32 (t=88)
c33 (t=136)
c34 (t=465)
c35 (t=450)
c36 (t=1214)
c37 (t=504)
c38 (t=50)
c39 (t=54)
c40 (t=801)
c41 (t=141)
c42 (t=718)
c43 (t=89)
c44 (t=169)
c45 (t=1461)
c46 (t=506)
c47 (t=32)
c48 (t=996)
Shared Genes
Shared Transcripts
1.0
-1.0
PCC

Fig. S16

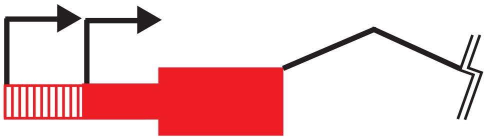
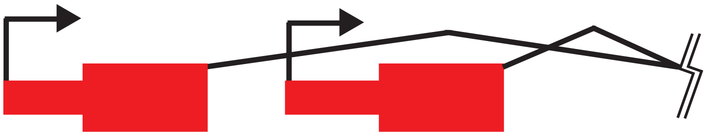
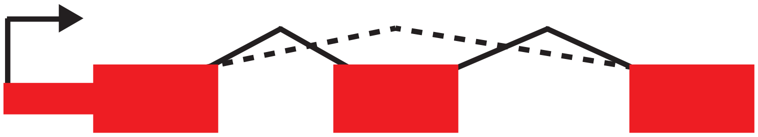
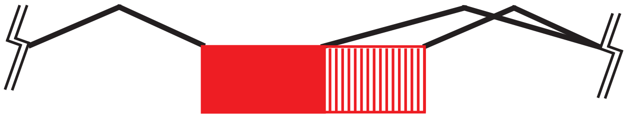
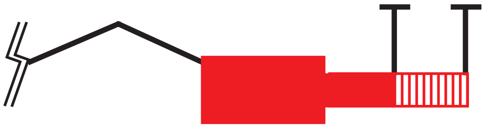
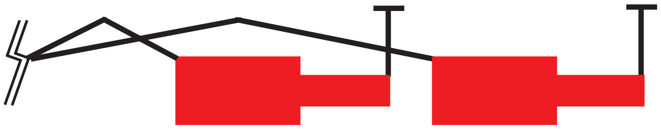
Type	# Genes	Transcript Model
Overlapping 5’ UTR	917	Multiple TSS Single ATG 
Distinct 5’ UTR	800	Multiple TSS Multiple ATG 
Alternative CDS Exon	1228	Alternate Exon 
Extended CDS Exon	1167	Extended Exon 
Overlapping 3’ UTR	1268	Multiple TTS Single STOP 
Distinct 3’ UTR	486	Multiple TTS Multiple STOP 

Fig. S17

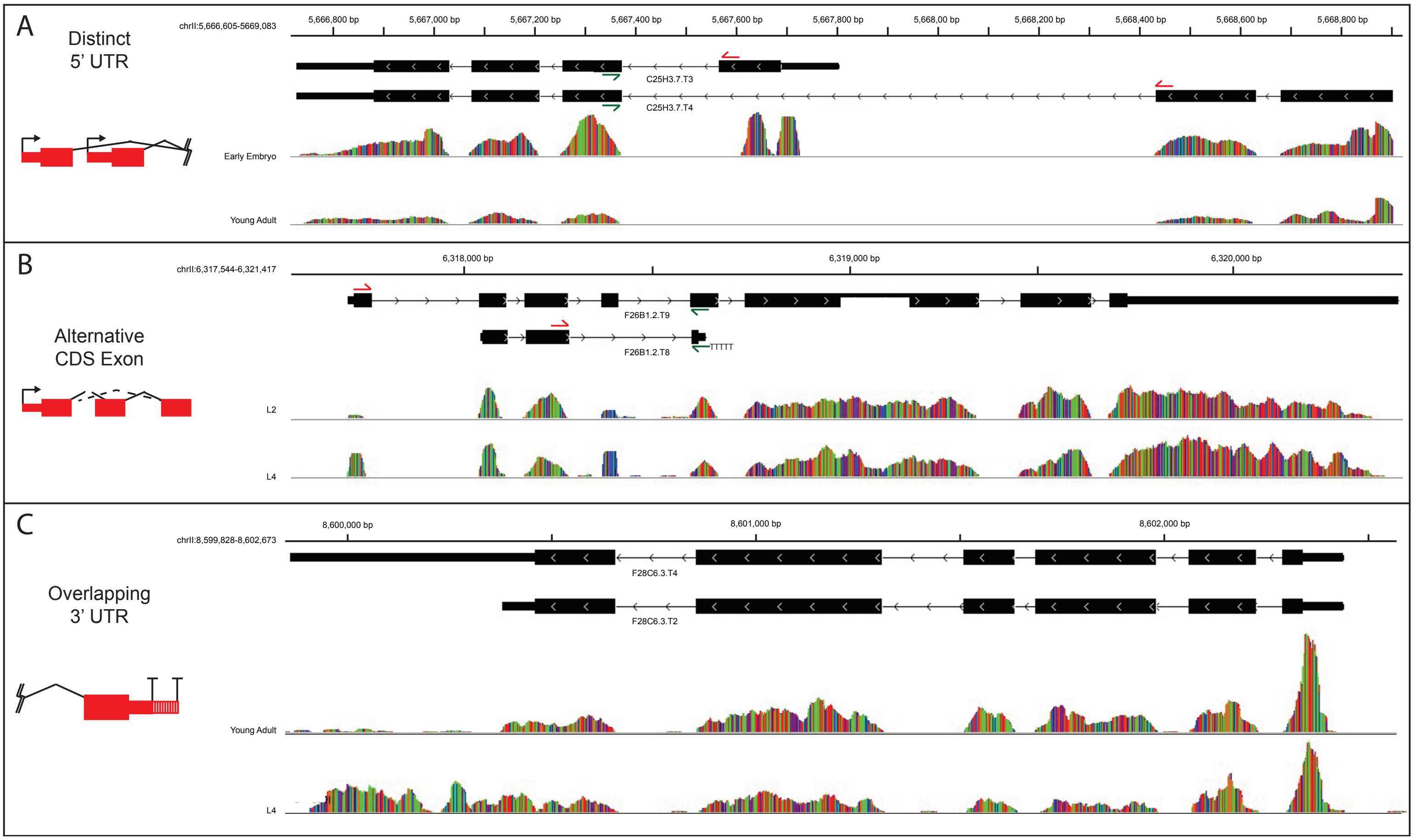

A
Distinct
5' UTR
chrII:5,666,605-5669,083
5,666,800 bp
5,667,000 bp
5,667,200 bp
5,667,400 bp
5,667,600 bp
5,667,800 bp
5,668,00 bp
5,668,200 bp
5,668,400 bp
5,668,600 bp
5,668,800 bp
C25H3.7.T3
C25H3.7.T4
Early Embryo
Young Adult
B
Alternative
CDS Exon
chrII:6,317,544-6,321,417
6,318,000 bp
6,319,000 bp
6,320,000 bp
F26B1.2.T9
F26B1.2.T8
TTTTT
L2
L4
C
Overlapping
3' UTR
chrII:8,599,828-8,602,673
8,600,000 bp
8,601,000 bp
8,602,000 bp
F28C6.3.T4
F28C6.3.T2
Young Adult
L4

Fig. S18

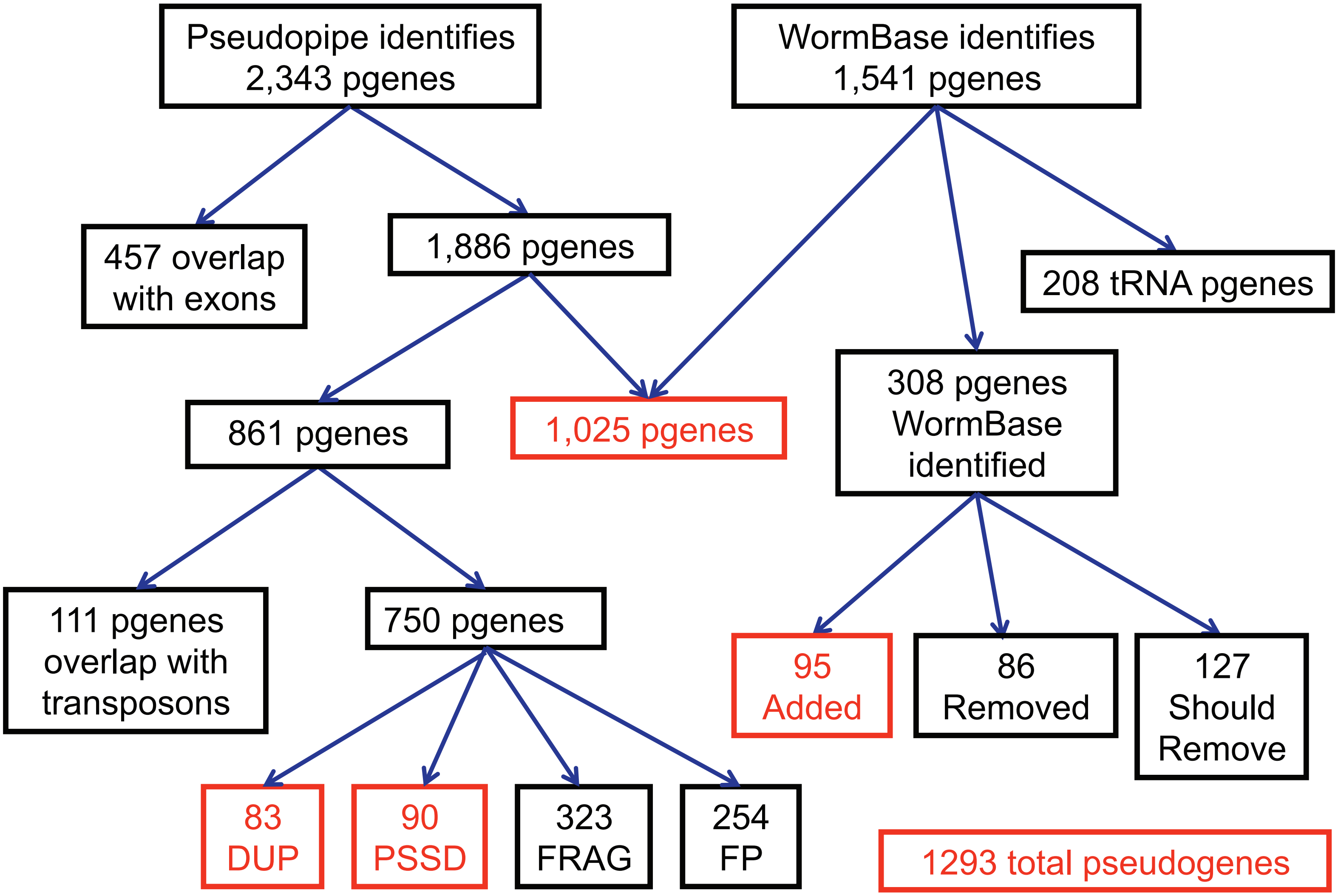

Pseudopipe identifies 2,343 pgenes
WormBase identifies 1,541 pgenes
457 overlap with exons
1,886 pgenes
208 tRNA pgenes
861 pgenes
1,025 pgenes
308 pgenes WormBase identified
111 pgenes overlap with transposons
750 pgenes
95 Added
86 Removed
127 Should Remove
83 DUP
90 PSSD
323 FRAG
254 FP
1293 total pseudogenes

Fig. S19

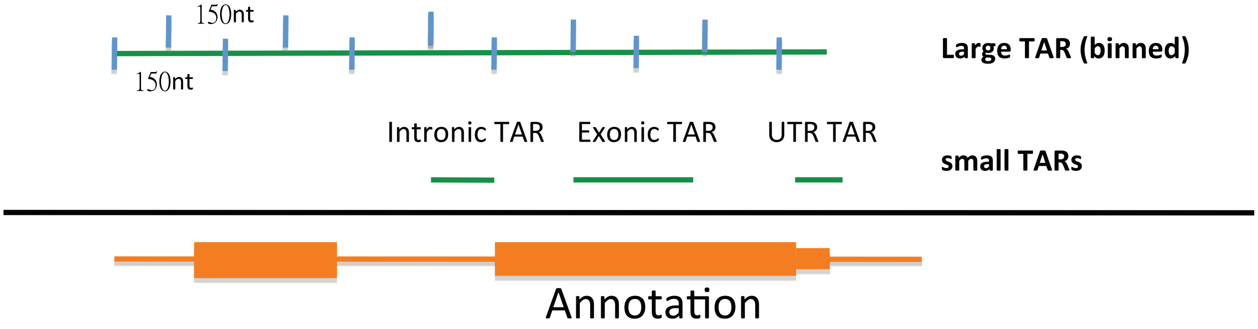

150nt
150nt
Large TAR (binned)
Intronic TAR
Exonic TAR
UTR TAR
small TARs
Annotation

Fig. S20

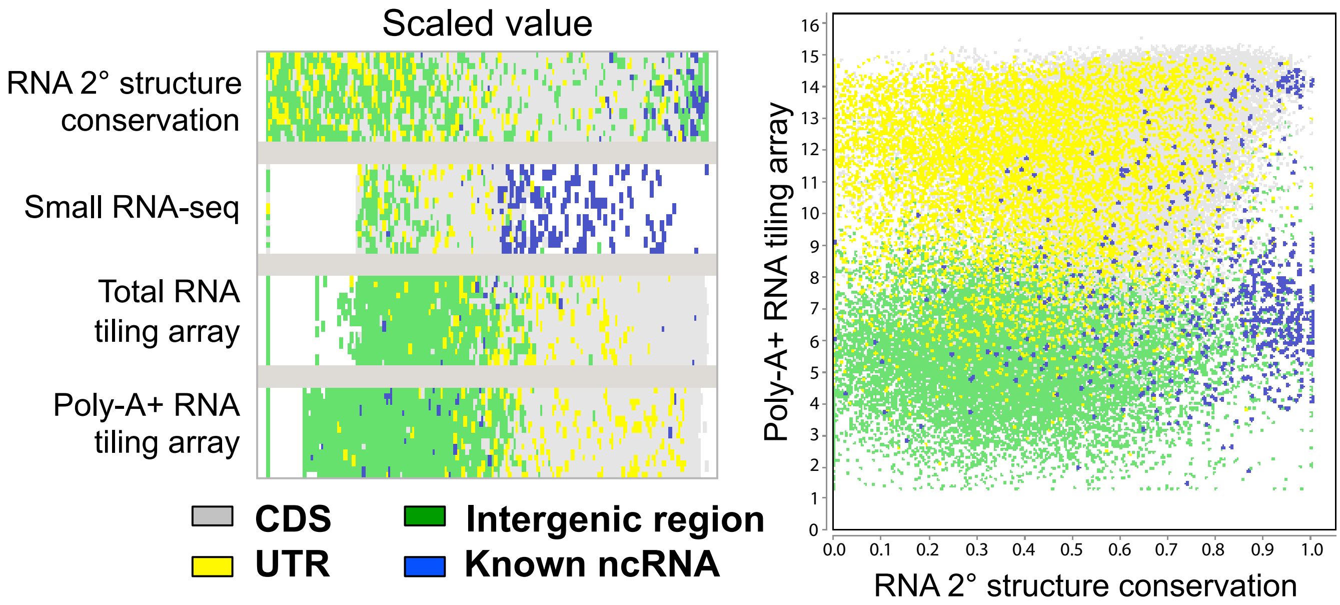

Scaled value
RNA 2° structure conservation
Small RNA-seq
Total RNA tiling array
Poly-A+ RNA tiling array
CDS
UTR
Intergenic region
Known ncRNA
Poly-A+ RNA tiling array
RNA 2° structure conservation
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Fig. S21A

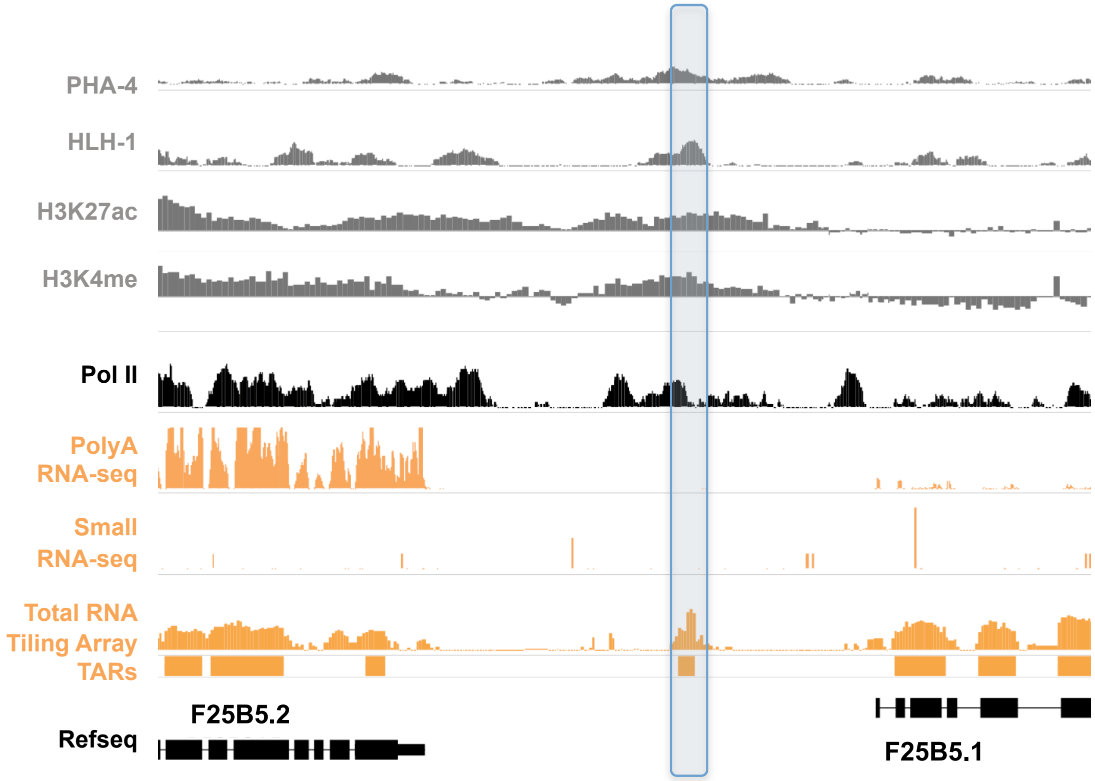

PHA-4
HLH-1
H3K27ac
H3K4me
Pol II
PolyA
RNA-seq
Small
RNA-seq
Total RNA
Tiling Array
TARs
F25B5.2
Refseq
F25B5.1

Fig. S21B

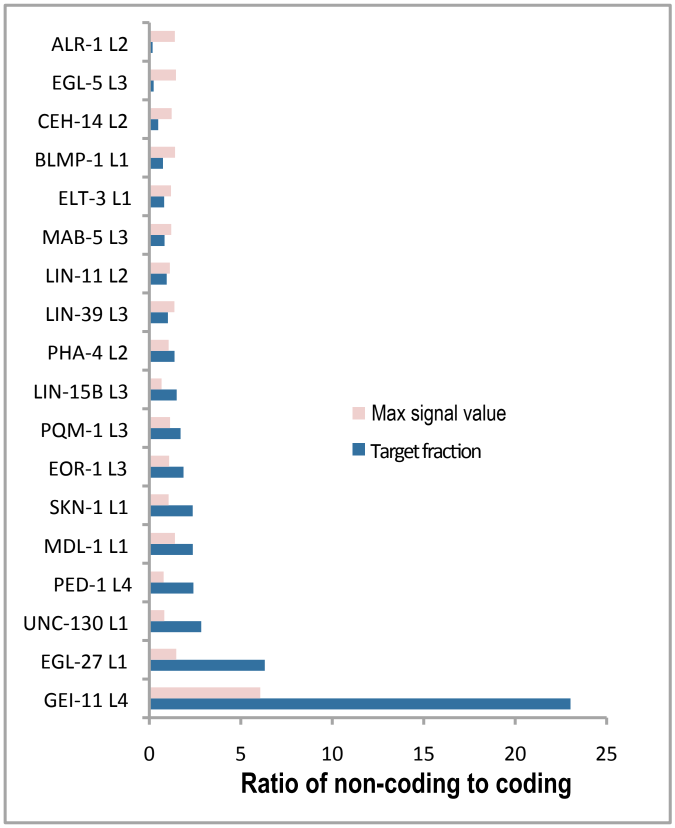

ALR-1 L2
EGL-5 L3
CEH-14 L2
BLMP-1 L1
ELT-3 L1
MAB-5 L3
LIN-11 L2
LIN-39 L3
PHA-4 L2
LIN-15B L3
PQM-1 L3
EOR-1 L3
SKN-1 L1
MDL-1 L1
PED-1 L4
UNC-130 L1
EGL-27 L1
GEI-11 L4
Max signal value
Target fraction
0
5
10
15
20
25
Ratio of non-coding to coding

Fig. S22A

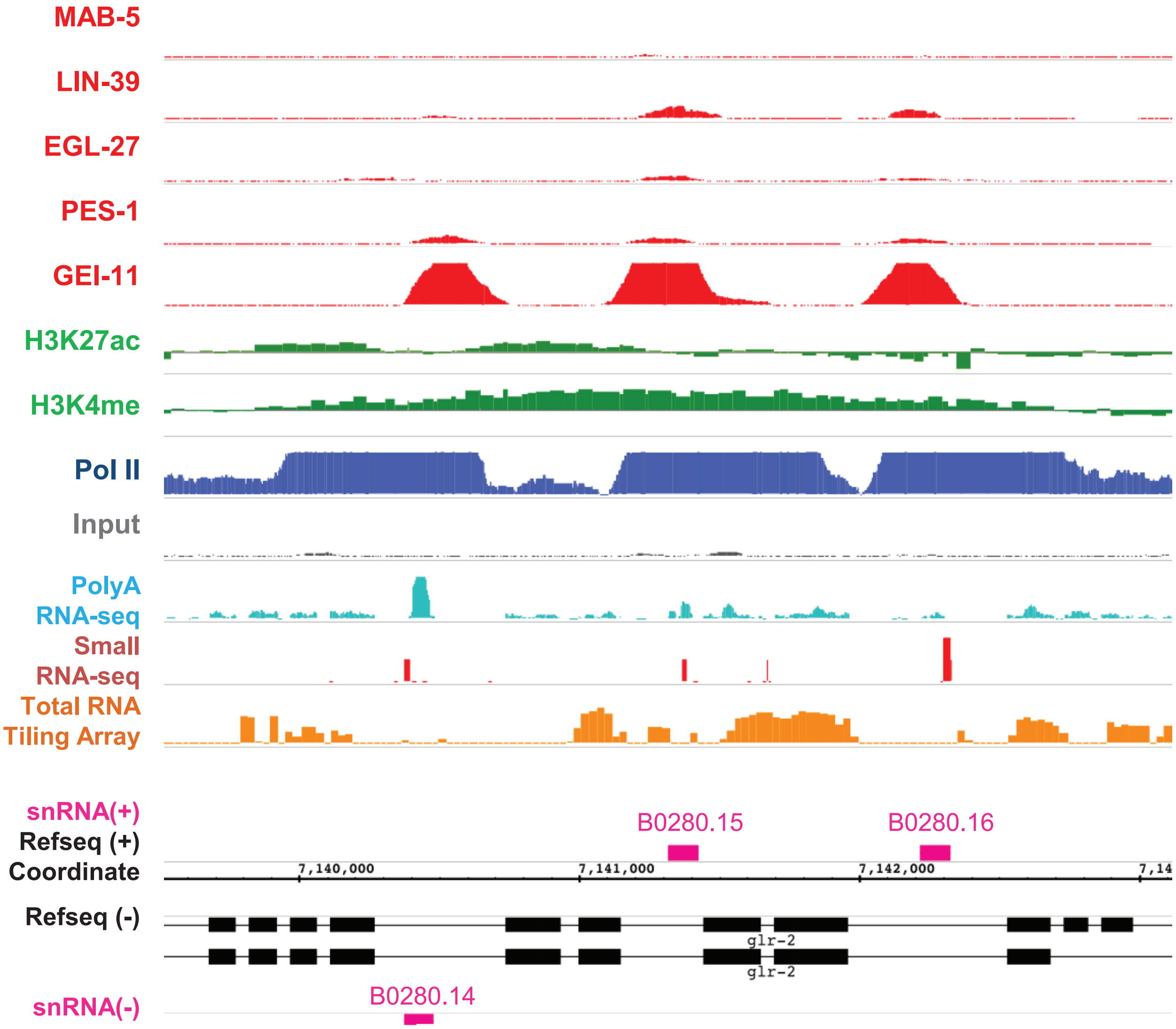

MAB-5
LIN-39
EGL-27
PES-1
GEI-11
H3K27ac
H3K4me
Pol II
Input
PolyA RNA-seq
Small RNA-seq
Total RNA Tiling Array
snRNA(+)
Refseq (+)
Coordinate
B0280.15
B0280.16
7,140,000
7,141,000
7,142,000
7,14
Refseq (-)
glr-2
glr-2
snRNA(-)
B0280.14

Fig. S22B

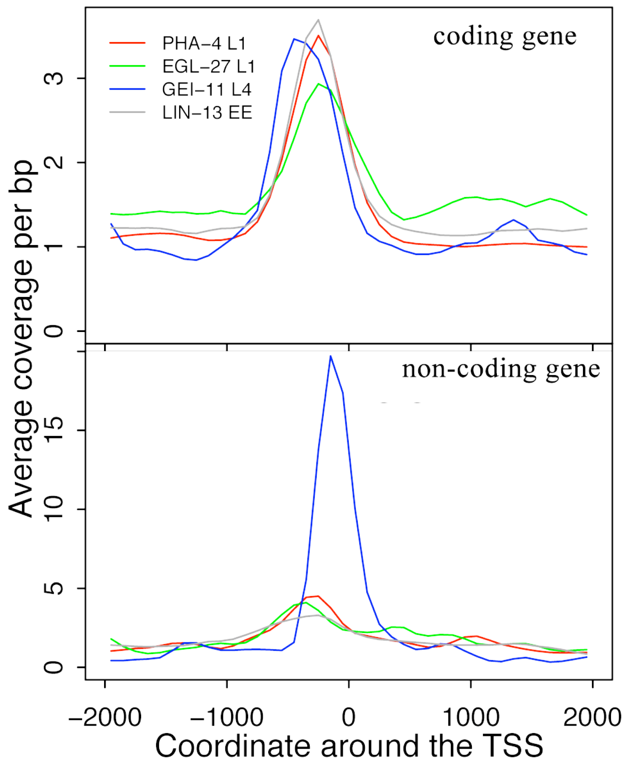

PHA−4 L1
EGL−27 L1
GEI−11 L4
LIN−13 EE
coding gene
non-coding gene
Average coverage per bp
0
1
2
3
0
5
10
15
−2000
−1000
0
1000
2000
Coordinate around the TSS

Fig. S22C

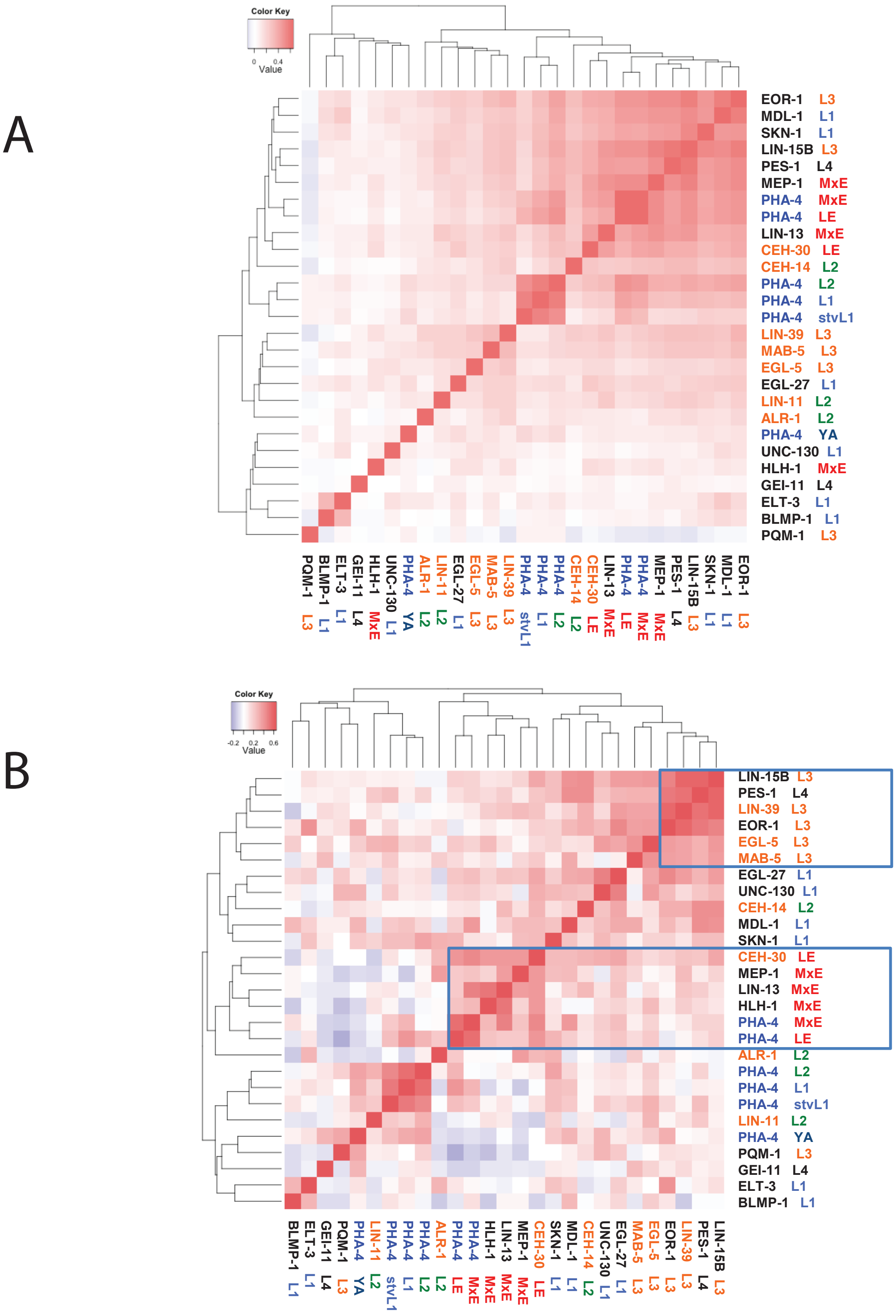
A
Color Key
0
0.4
Value
EOR-1 L3
MDL-1 L1
SKN-1 L1
LIN-15B L3
PES-1 L4
MEP-1 MxE
PHA-4 MxE
PHA-4 LE
LIN-13 MxE
CEH-30 LE
CEH-14 L2
PHA-4 L2
PHA-4 L1
PHA-4 stvL1
LIN-39 L3
MAB-5 L3
EGL-5 L3
EGL-27 L1
LIN-11 L2
ALR-1 L2
PHA-4 YA
UNC-130 L1
HLH-1 MxE
GEI-11 L4
ELT-3 L1
BLMP-1 L1
PQM-1 L3
PQM-1 L3
BLMP-1 L1
ELT-3 L1
GEI-11 L4
HLH-1 MxE
UNC-130 L1
PHA-4 YA
ALR-1 L2
LIN-11 L2
EGL-27 L1
EGL-5 L3
MAB-5 L3
LIN-39 L3
PHA-4 stvL1
PHA-4 L1
PHA-4 L2
CEH-14 L2
CEH-30 LE
LIN-13 MxE
PHA-4 LE
PHA-4 MxE
MEP-1 MxE
PES-1 L4
LIN-15B L3
SKN-1 L1
MDL-1 L1
EOR-1 L3
B
Color Key
-0.2
0.2
0.6
Value
LIN-15B L3
PES-1 L4
LIN-39 L3
EOR-1 L3
EGL-5 L3
MAB-5 L3
EGL-27 L1
UNC-130 L1
CEH-14 L2
MDL-1 L1
SKN-1 L1
CEH-30 LE
MEP-1 MxE
LIN-13 MxE
HLH-1 MxE
PHA-4 MxE
PHA-4 LE
ALR-1 L2
PHA-4 L2
PHA-4 L1
PHA-4 stvL1
LIN-11 L2
PHA-4 YA
PQM-1 L3
GEI-11 L4
ELT-3 L1
BLMP-1 L1
BLMP-1 L1
ELT-3 L1
GEI-11 L4
PQM-1 L3
PHA-4 YA
LIN-11 L2
PHA-4 stvL1
PHA-4 L1
PHA-4 L2
ALR-1 L2
PHA-4 LE
PHA-4 MxE
HLH-1 MxE
LIN-13 MxE
MEP-1 MxE
CEH-30 LE
SKN-1 L1
MDL-1 L1
CEH-14 L2
UNC-130 L1
EGL-27 L1
MAB-5 L3
EGL-5 L3
EOR-1 L3
LIN-39 L3
PES-1 L4
LIN-15B L3

Fig. S23

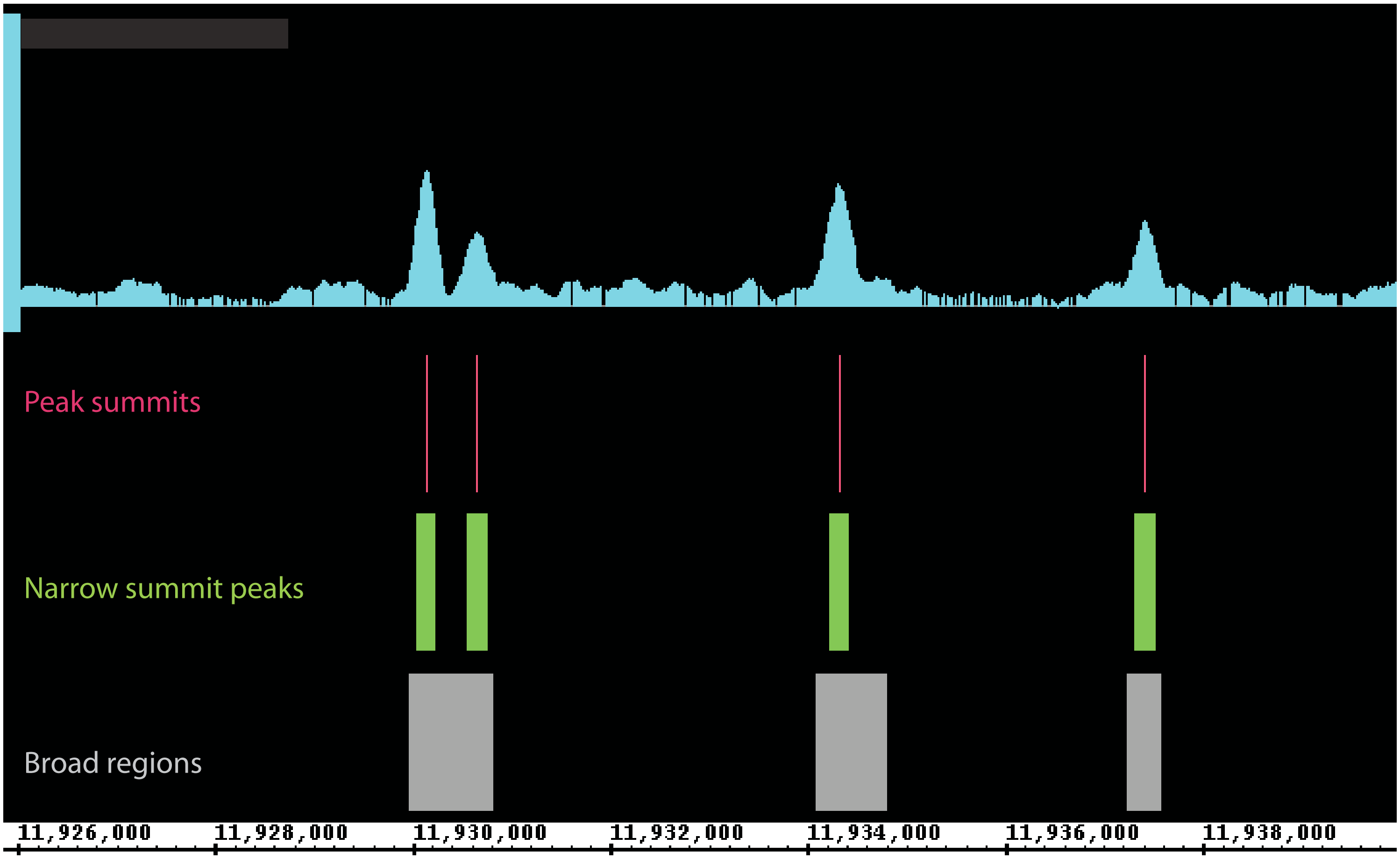
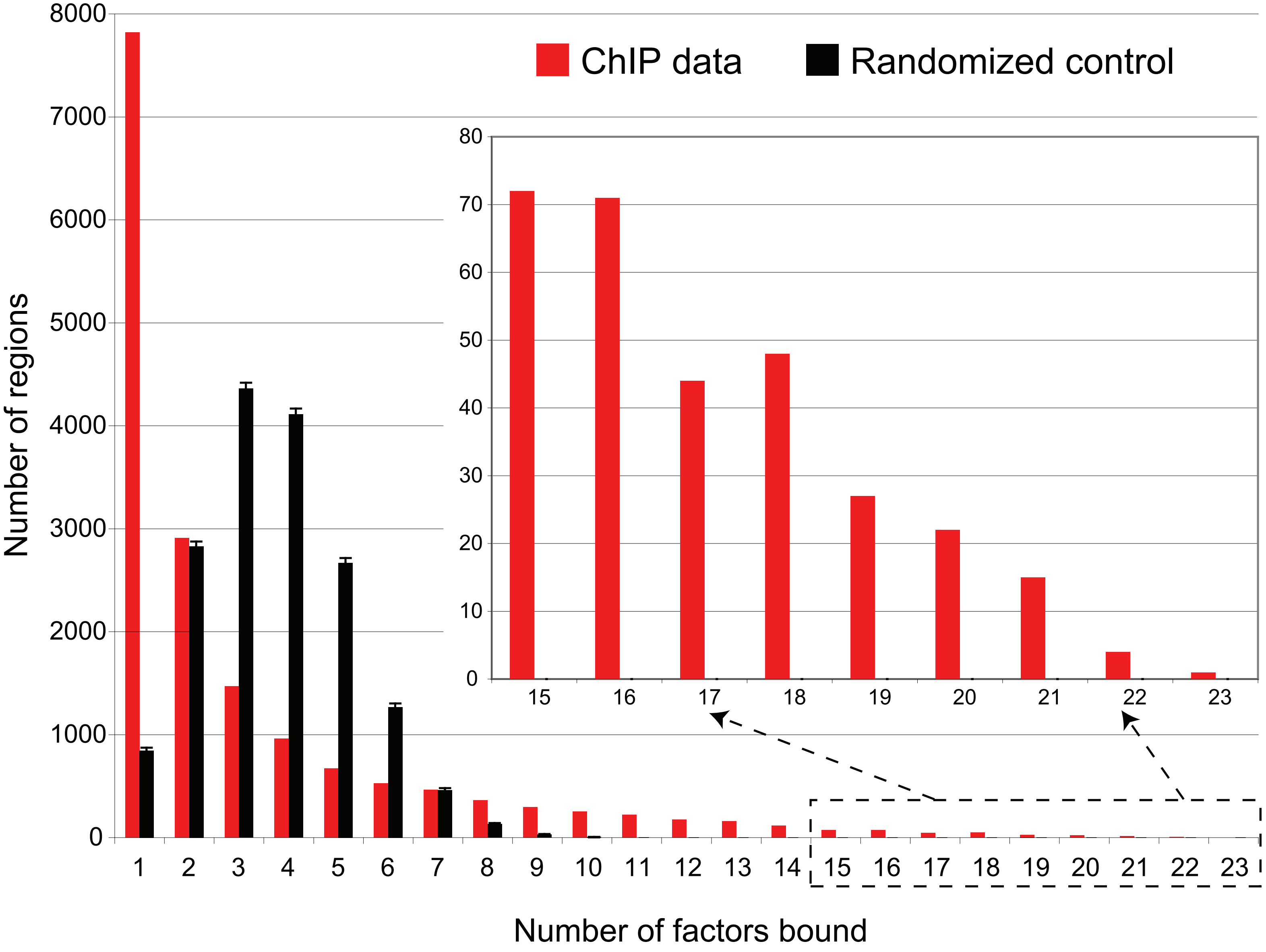

Peak summits
Narrow summit peaks
Broad regions
11,926,000
11,928,000
11,930,000
11,932,000
11,934,000
11,936,000
11,938,000

Fig. S24



8000
7000
6000
5000
4000
3000
2000
1000
0
Number of regions
🟥 ChIP data
■ Randomized control
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Number of factors bound
80
70
60
50
40
30
20
10
0
15
16
17
18
19
20
21
22
23

Fig. S25A

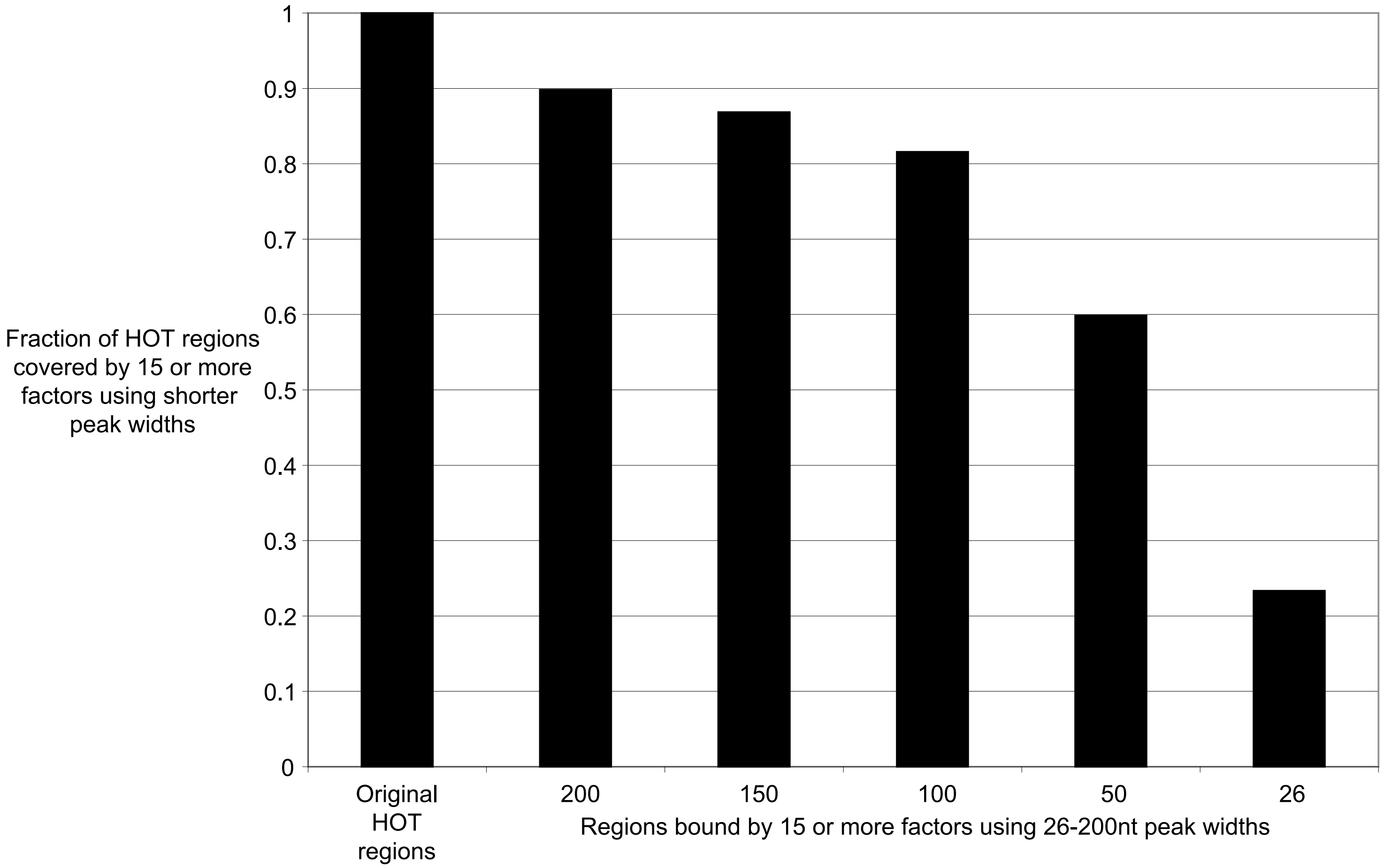

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Fraction of HOT regions covered by 15 or more factors using shorter peak widths
Original HOT regions
200
150
100
50
26
Regions bound by 15 or more factors using 26-200nt peak widths

Fig. S25B

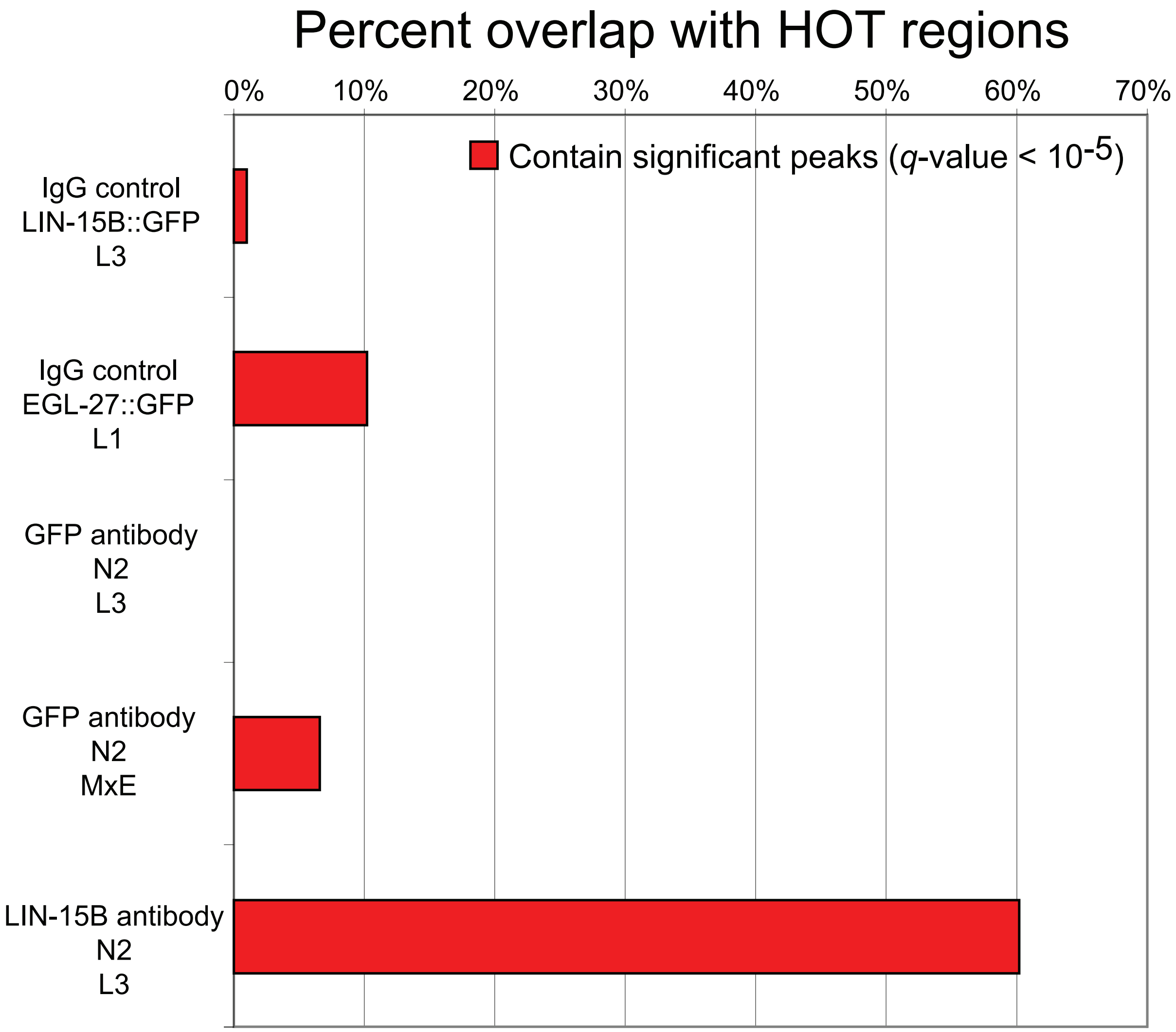

Percent overlap with HOT regions
0%
10%
20%
30%
40%
50%
60%
70%
🟥 Contain significant peaks (q-value < 10^{-5})
IgG control
LIN-15B::GFP
L3
IgG control
EGL-27::GFP
L1
GFP antibody
N2
L3
GFP antibody
N2
MxE
LIN-15B antibody
N2
L3

Fig. S26A

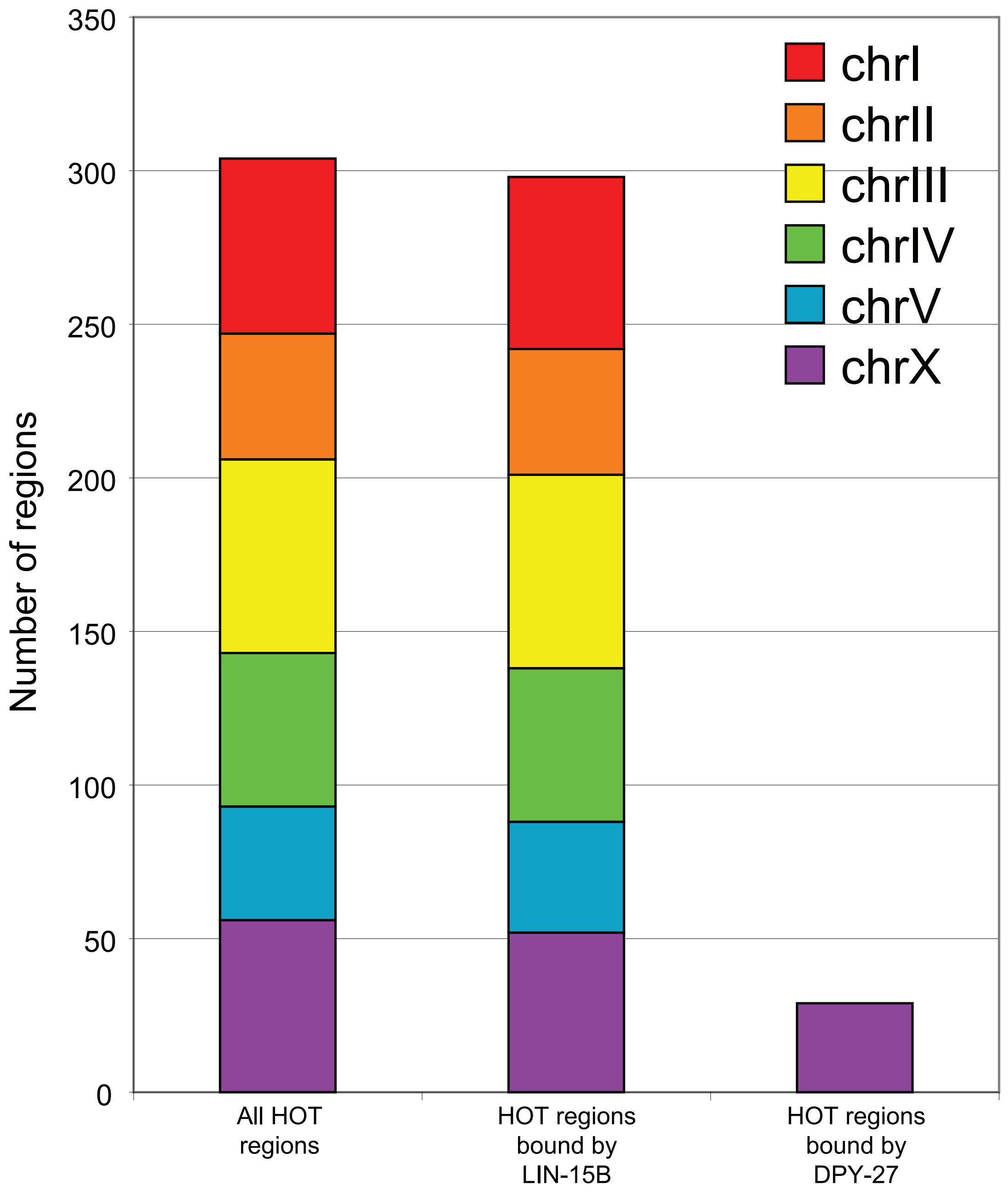

350
300
250
200
150
100
50
0
Number of regions
chrI
chrII
chrIII
chrIV
chrV
chrX
All HOT regions
HOT regions bound by LIN-15B
HOT regions bound by DPY-27

Fig. S26B

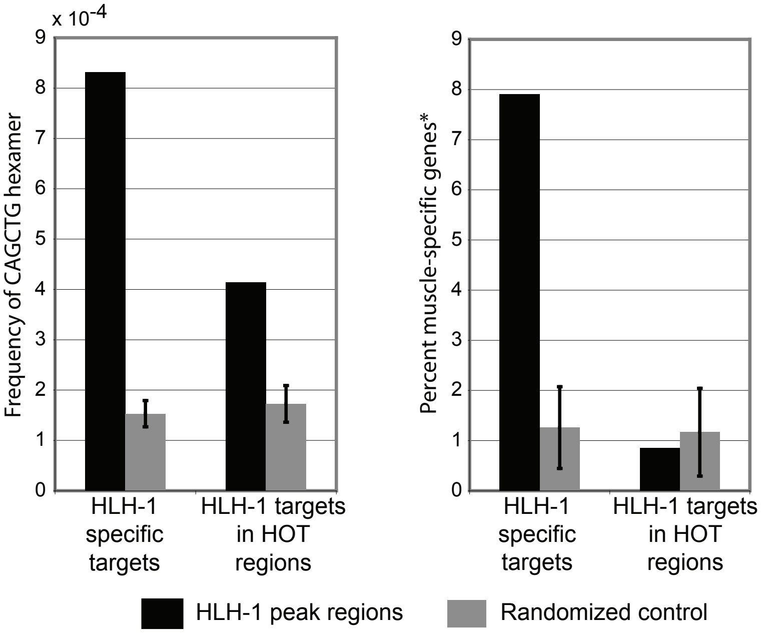

x 10^-4
Frequency of CAGCTG hexamer
9
8
7
6
5
4
3
2
1
0
HLH-1 specific targets
HLH-1 targets in HOT regions
Percent muscle-specific genes*
9
8
7
6
5
4
3
2
1
0
HLH-1 specific targets
HLH-1 targets in HOT regions
HLH-1 peak regions
Randomized control

Fig. S27A

Factor-specific targets compared to HOT targets

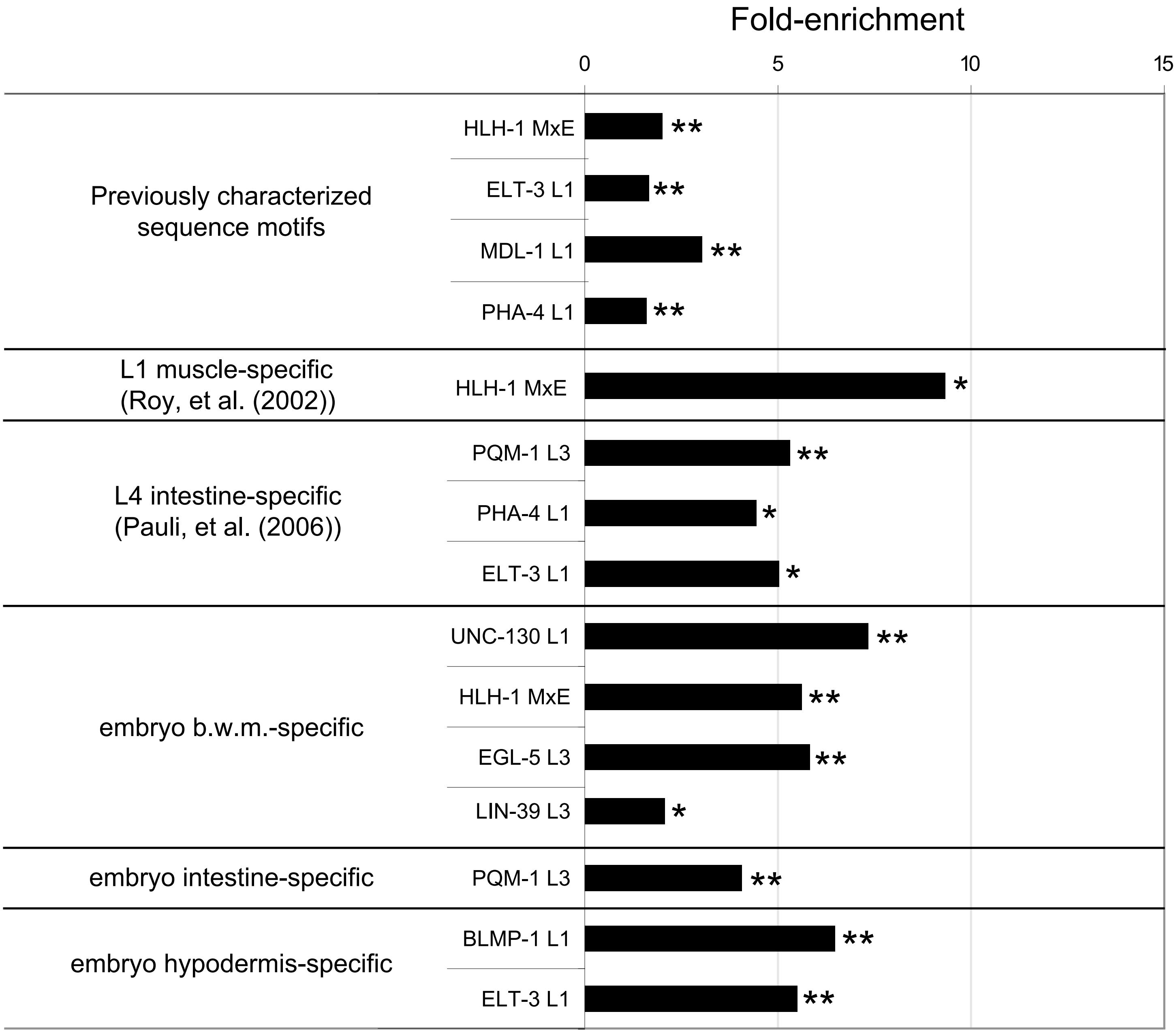


Fig. S27B

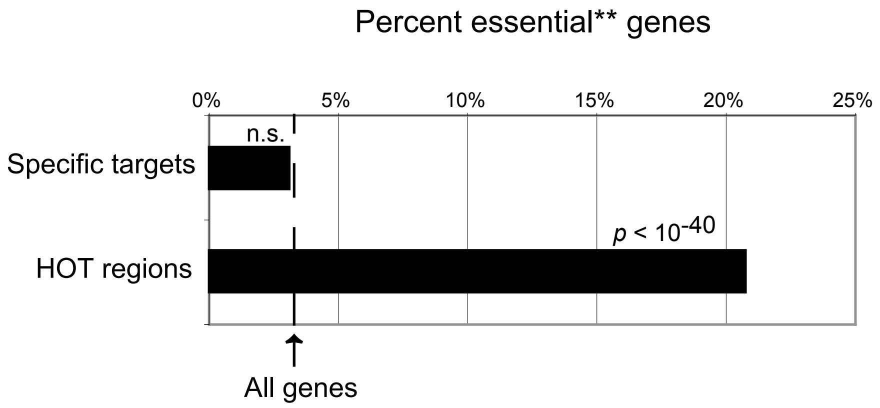

Percent essential** genes
0%
5%
10%
15%
20%
25%
n.s.
Specific targets
p < 10^{-40}
HOT regions
All genes

Fig. S27C

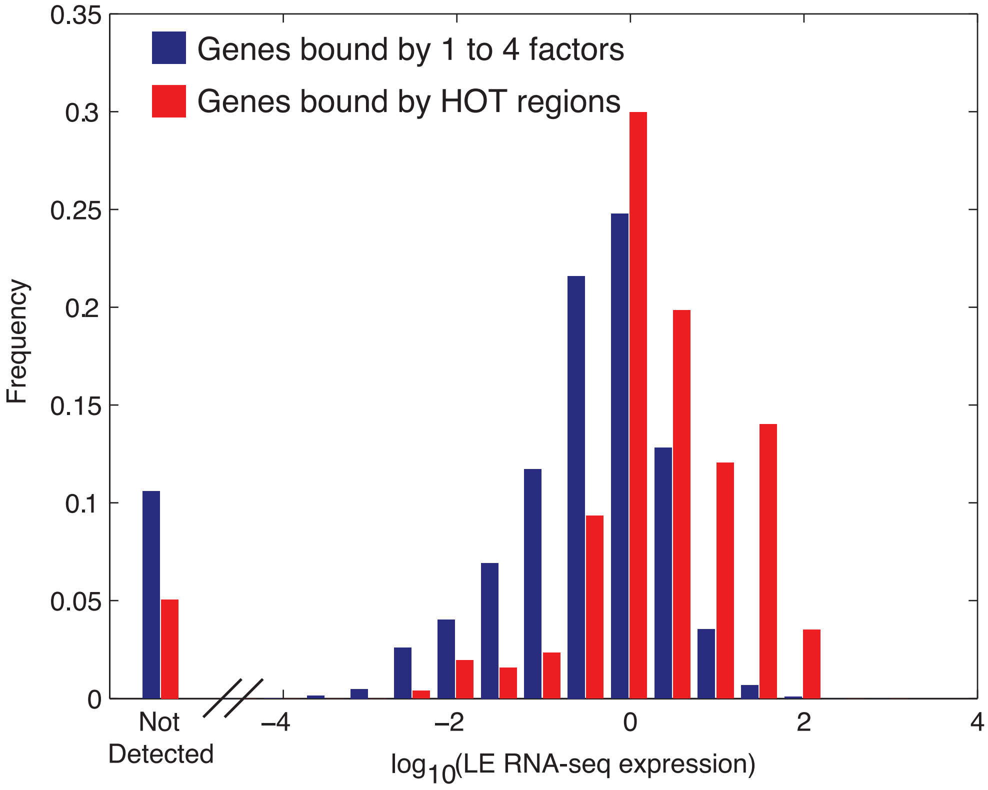

0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Frequency
Genes bound by 1 to 4 factors
Genes bound by HOT regions
Not Detected
−4
−2
0
2
4
$\log_{10}$(LE RNA-seq expression)

Fig. S28A

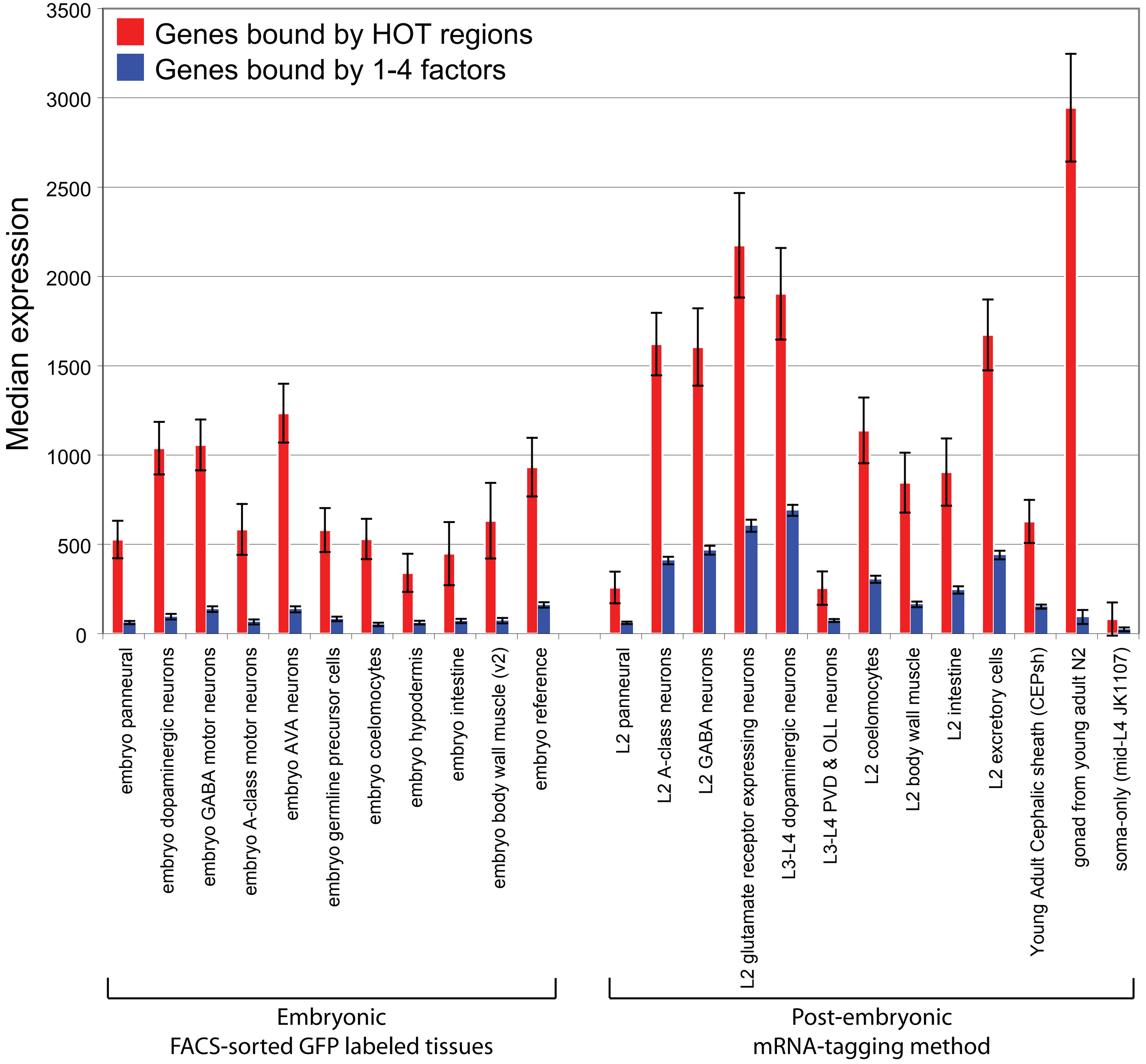
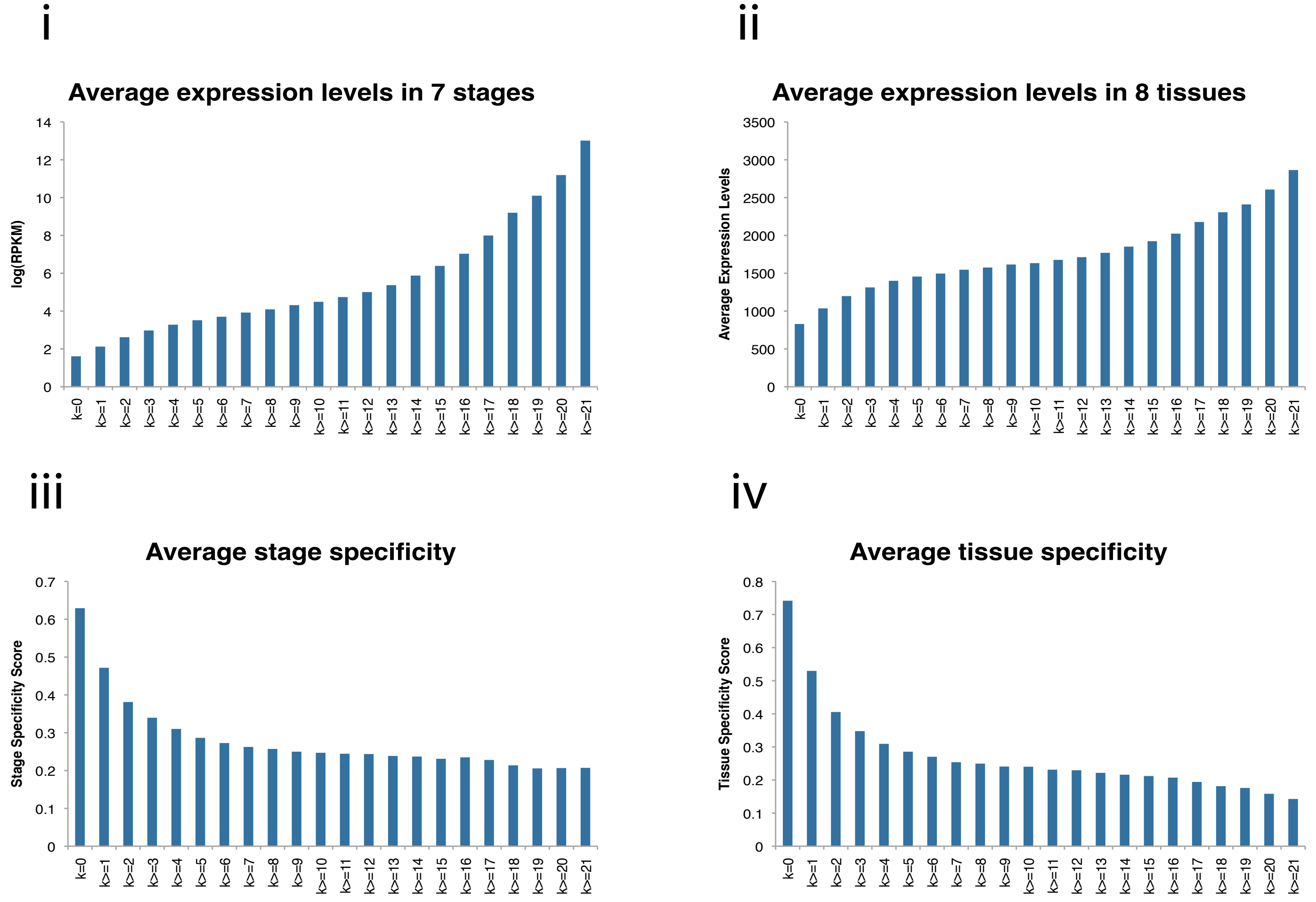

Genes bound by HOT regions
Genes bound by 1-4 factors
Median expression
3500
3000
2500
2000
1500
1000
500
0
embryo panneural
embryo dopaminergic neurons
embryo GABA motor neurons
embryo A-class motor neurons
embryo AVA neurons
embryo germline precursor cells
embryo coelomocytes
embryo hypodermis
embryo intestine
embryo body wall muscle (v2)
embryo reference
L2 panneural
L2 A-class neurons
L2 GABA neurons
L2 glutamate receptor expressing neurons
L3-L4 dopaminergic neurons
L3-L4 PVD & OLL neurons
L2 coelomocytes
L2 body wall muscle
L2 intestine
L2 excretory cells
Young Adult Cephalic sheath (CEPsh)
gonad from young adult N2
soma-only (mid-L4 JK1107)
Embryonic
FACS-sorted GFP labeled tissues
Post-embryonic
mRNA-tagging method

Fig. S28B



i
Average expression levels in 7 stages
log(RPKM)
0
2
4
6
8
10
12
14
k=0
k>=1
k>=2
k>=3
k>=4
k>=5
k>=6
k>=7
k>=8
k>=9
k>=10
k>=11
k>=12
k>=13
k>=14
k>=15
k>=16
k>=17
k>=18
k>=19
k>=20
k>=21
ii
Average expression levels in 8 tissues
Average Expression Levels
0
500
1000
1500
2000
2500
3000
3500
k=0
k>=1
k>=2
k>=3
k>=4
k>=5
k>=6
k>=7
k>=8
k>=9
k>=10
k>=11
k>=12
k>=13
k>=14
k>=15
k>=16
k>=17
k>=18
k>=19
k>=20
k>=21
iii
Average stage specificity
Stage Specificity Score
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
k=0
k>=1
k>=2
k>=3
k>=4
k>=5
k>=6
k>=7
k>=8
k>=9
k>=10
k>=11
k>=12
k>=13
k>=14
k>=15
k>=16
k>=17
k>=18
k>=19
k>=20
k>=21
iv
Average tissue specificity
Tissue Specificity Score
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
k=0
k>=1
k>=2
k>=3
k>=4
k>=5
k>=6
k>=7
k>=8
k>=9
k>=10
k>=11
k>=12
k>=13
k>=14
k>=15
k>=16
k>=17
k>=18
k>=19
k>=20
k>=21

Fig. S28C

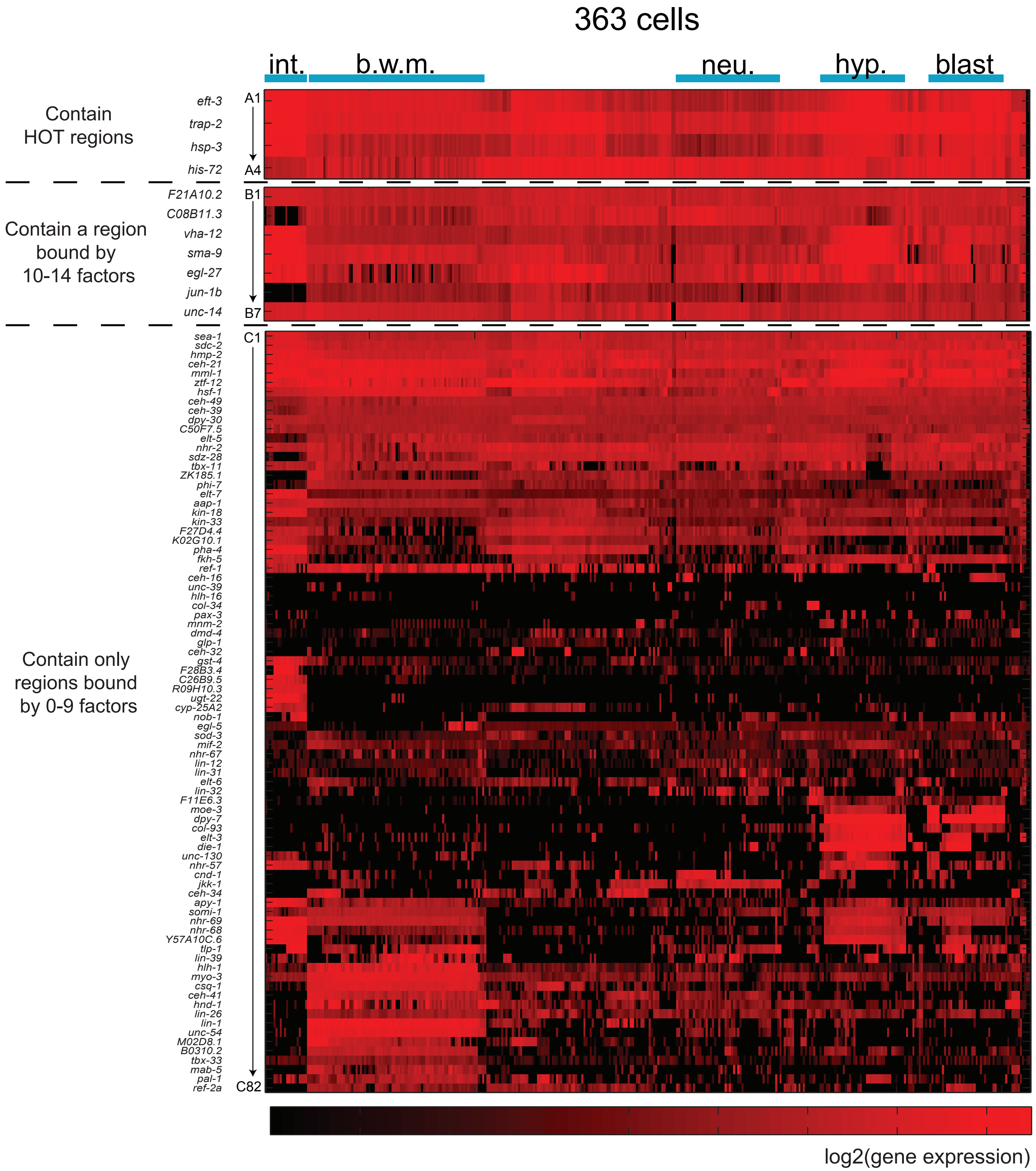

363 cells
int.
b.w.m.
neu.
hyp.
blast
Contain HOT regions
eft-3
trap-2
hsp-3
his-72
A1
A4
Contain a region bound by 10-14 factors
F21A10.2
C08B11.3
vha-12
sma-9
egl-27
jun-1b
unc-14
B1
B7
Contain only regions bound by 0-9 factors
sea-1
sdc-2
hmp-2
ceh-21
mml-1
ztf-12
hsf-1
ceh-49
ceh-39
dpy-30
C50F7.5
elt-5
nhr-2
sdz-28
tbx-11
ZK185.1
phi-7
elt-7
aap-1
kin-18
kin-33
F27D4.4
K02G10.1
pha-4
fkh-5
ref-1
ceh-16
unc-39
hlh-16
col-34
pax-3
mnm-2
dmd-4
glp-1
ceh-32
gst-4
F28B3.4
C26B9.5
R09H10.3
ugt-22
cyp-25A2
nob-1
egl-5
sod-3
mif-2
nhr-67
lin-12
lin-31
elt-6
lin-32
F11E6.3
moe-3
dpy-7
col-93
elt-3
die-1
unc-130
nhr-57
cnd-1
jkk-1
ceh-34
apy-1
somi-1
nhr-69
nhr-68
Y57A10C.6
tlp-1
lin-39
hlh-1
myo-3
csq-1
ceh-41
hnd-1
lin-26
lin-1
unc-54
M02D8.1
B0310.2
tbx-33
mab-5
pal-1
ref-2a
C1
C82
log2(gene expression)

Fig. S29

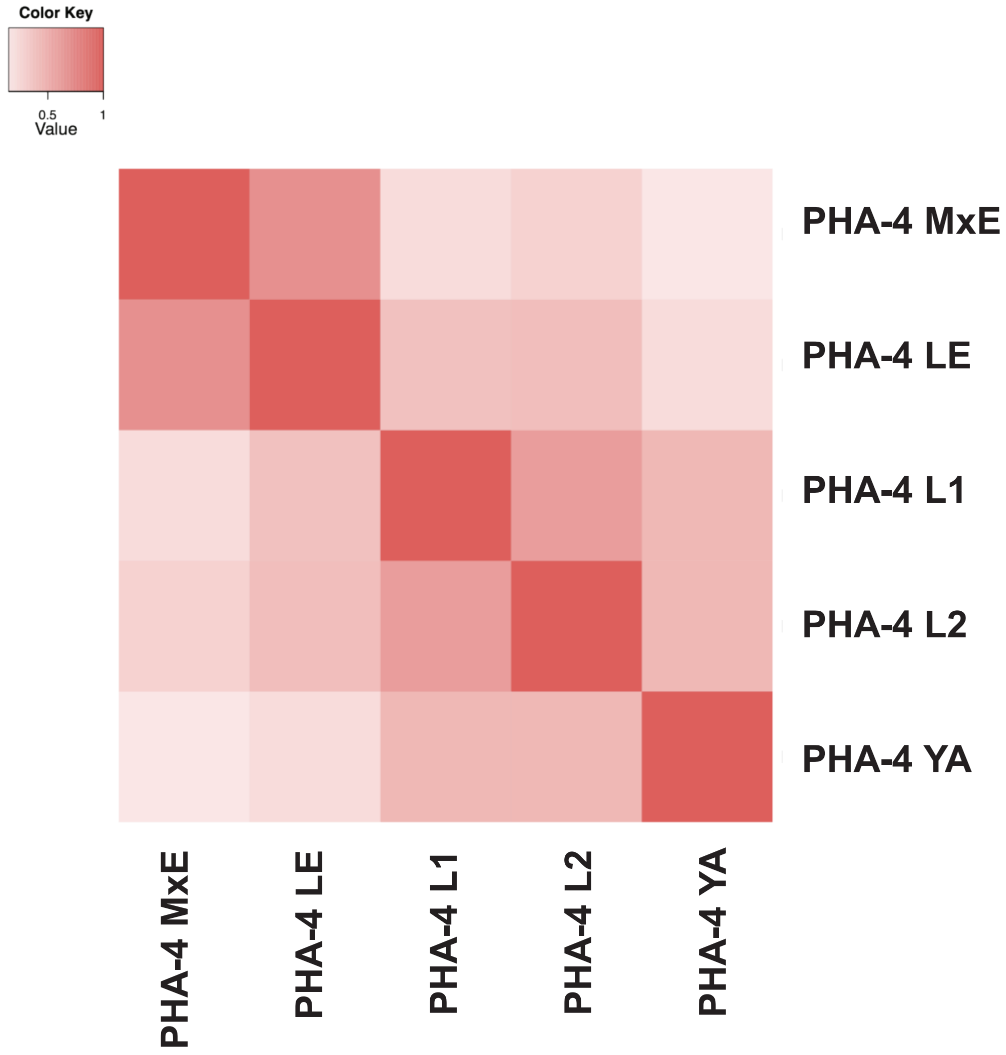

Color Key
0.5
1
Value
PHA-4 MxE
PHA-4 LE
PHA-4 L1
PHA-4 L2
PHA-4 YA
PHA-4 MxE
PHA-4 LE
PHA-4 L1
PHA-4 L2
PHA-4 YA

Fig. S30

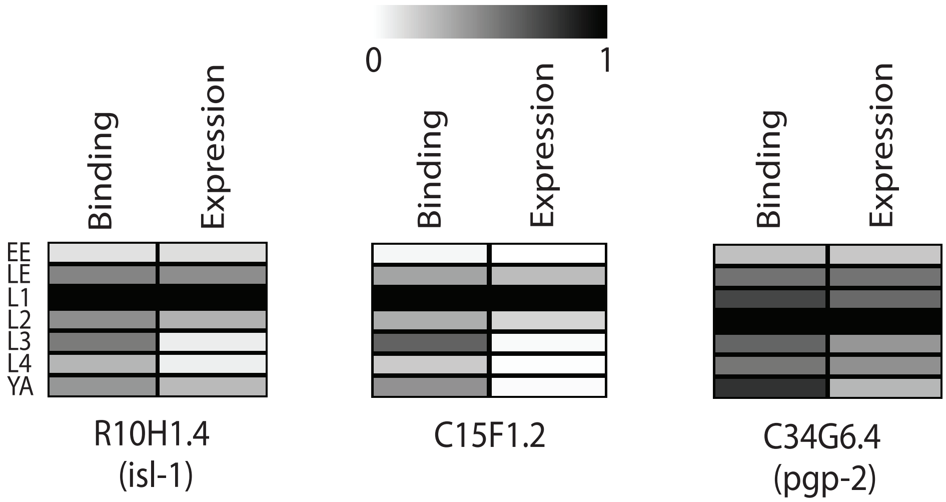
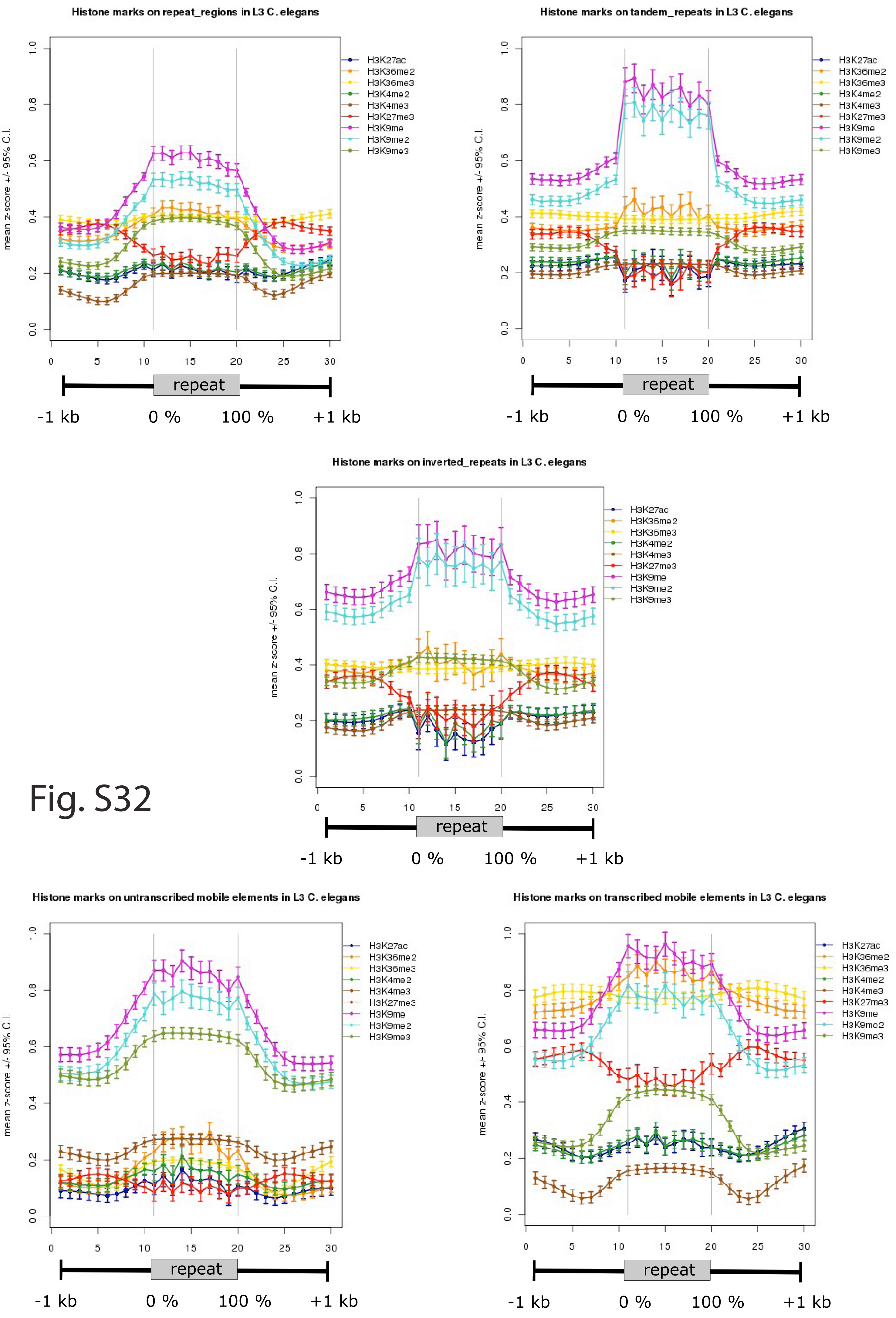

0
1
Binding
Expression
EE
LE
L1
L2
L3
L4
YA
R10H1.4
(isl-1)
C15F1.2
C34G6.4
(pgp-2)

Fig. S31

Fig. S32



Histone marks on repeat_regions in L3 C. elegans

Histone marks on tandem_repeats in L3 C. elegans

Histone marks on inverted_repeats in L3 C. elegans

Histone marks on untranscribed mobile elements in L3 C. elegans

Histone marks on transcribed mobile elements in L3 C. elegans

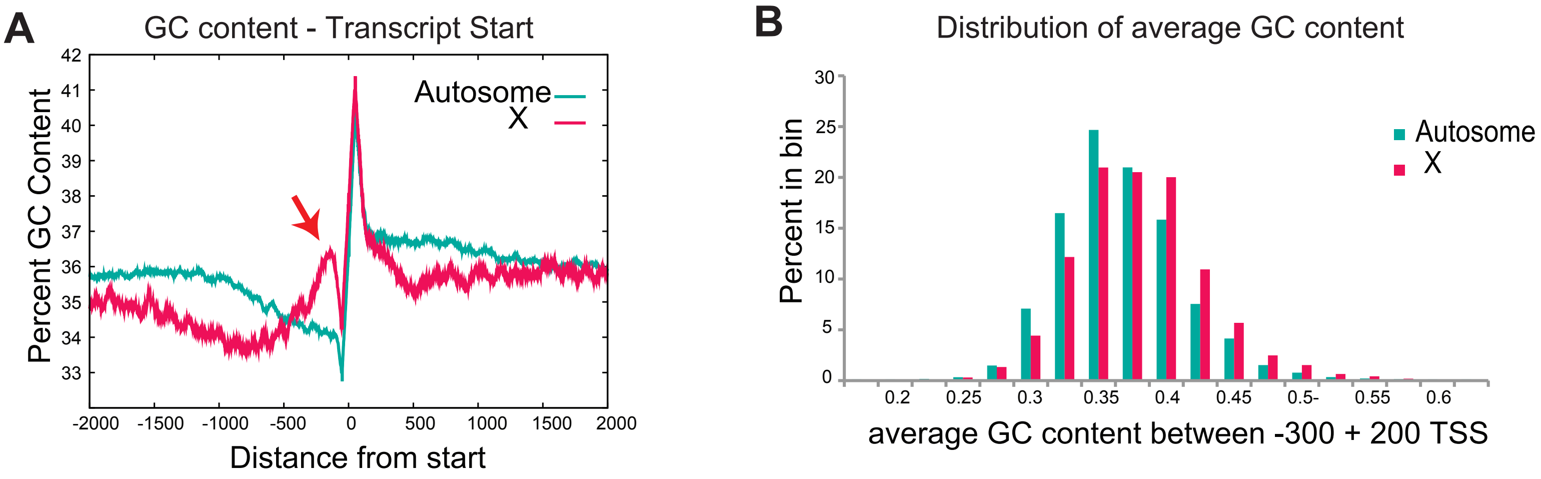

A
GC content - Transcript Start
Percent GC Content
42
41
40
39
38
37
36
35
34
33
Autosome
X
-2000
-1500
-1000
-500
0
500
1000
1500
2000
Distance from start
B
Distribution of average GC content
Percent in bin
30
25
20
15
10
5
0
🟩 Autosome
🟥 X
0.2
0.25
0.3
0.35
0.4
0.45
0.5-
0.55
0.6
average GC content between -300 + 200 TSS

Fig. S33

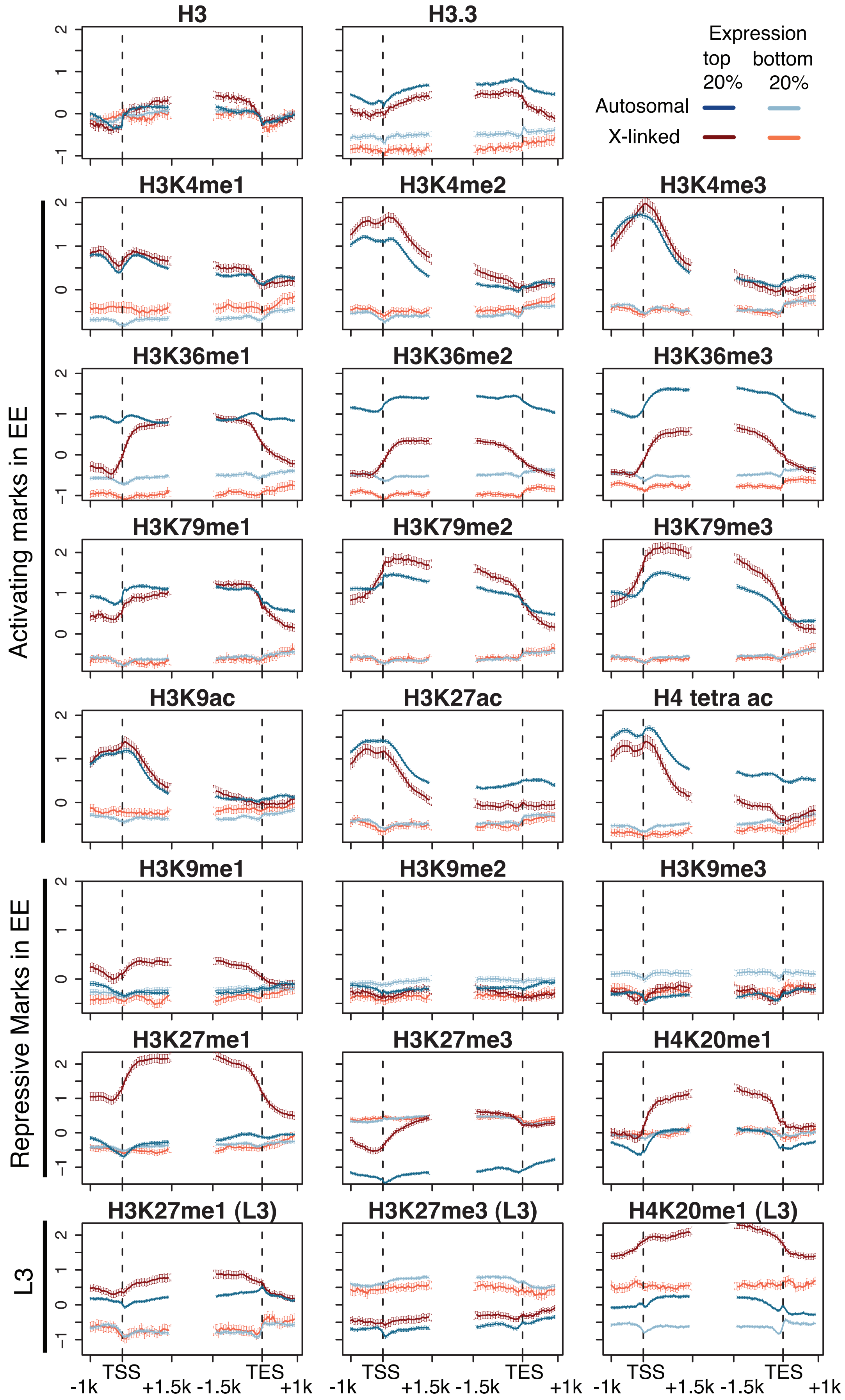

H3
H3.3
Expression
top 20%
bottom 20%
Autosomal
X-linked
H3K4me1
H3K4me2
H3K4me3
H3K36me1
H3K36me2
H3K36me3
H3K79me1
H3K79me2
H3K79me3
H3K9ac
H3K27ac
H4 tetra ac
Activating marks in EE
H3K9me1
H3K9me2
H3K9me3
H3K27me1
H3K27me3
H4K20me1
Repressive Marks in EE
H3K27me1 (L3)
H3K27me3 (L3)
H4K20me1 (L3)
L3
2
1
0
-1
-1k
TSS
+1.5k
-1.5k
TES
+1k

Fig. S34

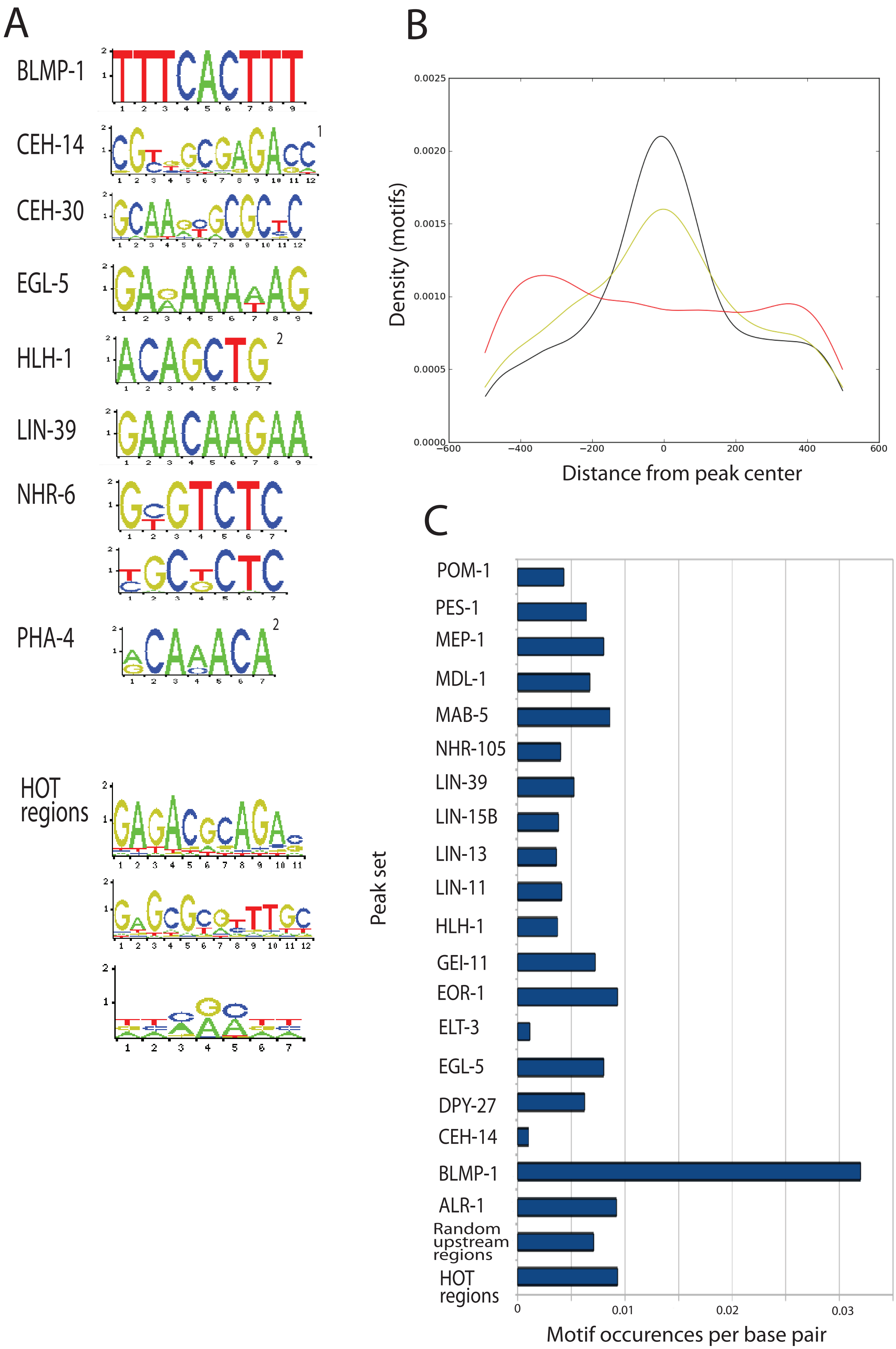

A
BLMP-1
CEH-14
CEH-30
EGL-5
HLH-1
LIN-39
NHR-6
PHA-4
HOT regions
B
Density (motifs)
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
−600
−400
−200
0
200
400
600
Distance from peak center
C
Peak set
POM-1
PES-1
MEP-1
MDL-1
MAB-5
NHR-105
LIN-39
LIN-15B
LIN-13
LIN-11
HLH-1
GEI-11
EOR-1
ELT-3
EGL-5
DPY-27
CEH-14
BLMP-1
ALR-1
Random upstream regions
HOT regions
0
0.01
0.02
0.03
Motif occurences per base pair

Fig. S35

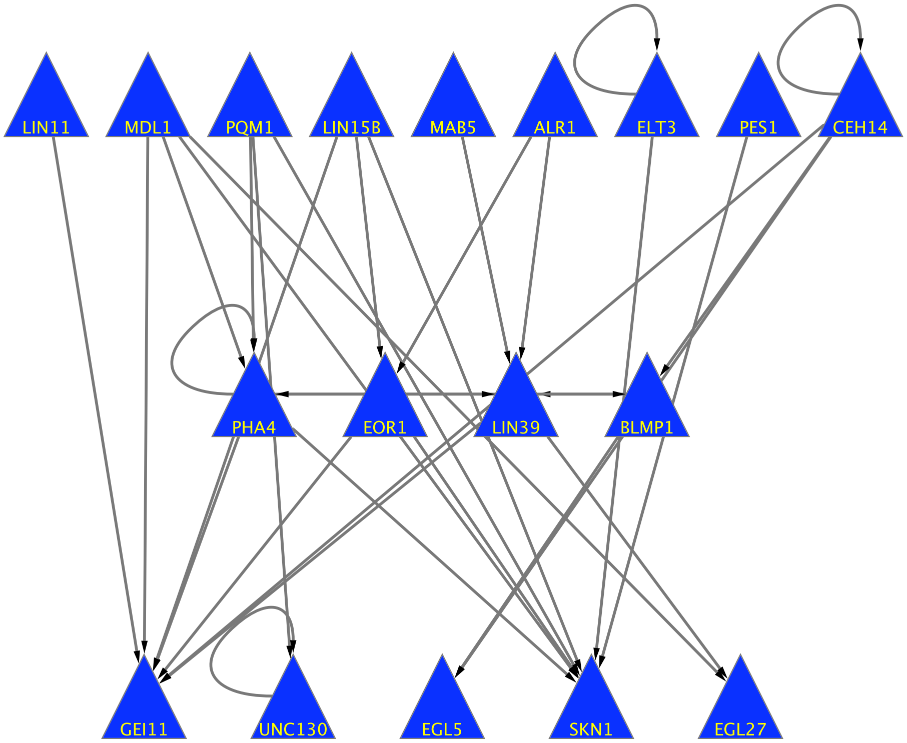

LIN11
MDL1
PQM1
LIN15B
MAB5
ALR1
ELT3
PES1
CEH14
PHA4
EOR1
LIN39
BLMP1
GEI11
UNC130
EGL5
SKN1
EGL27

Fig. S36

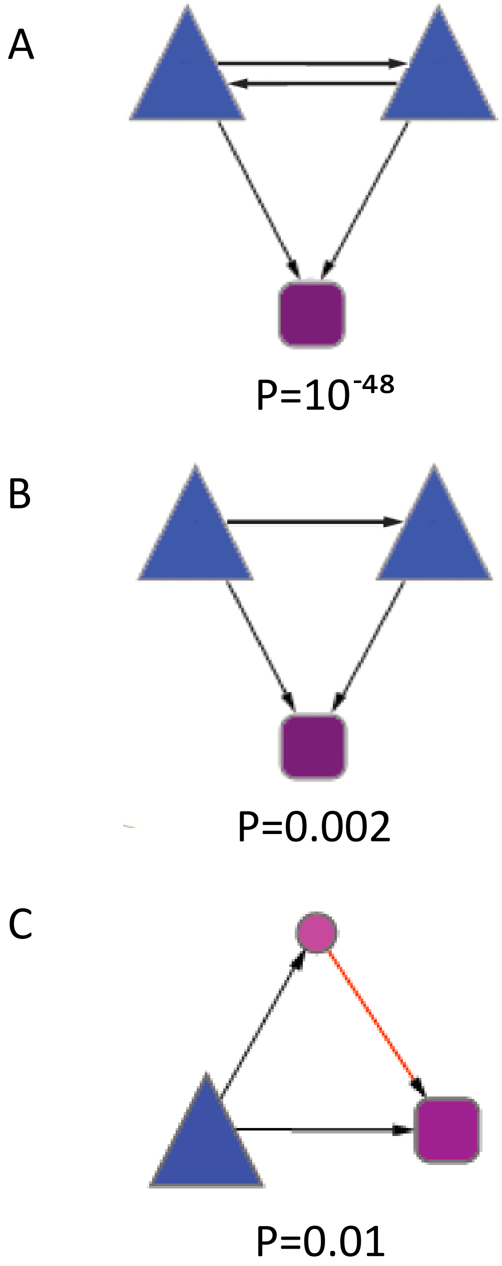


Fig. S37

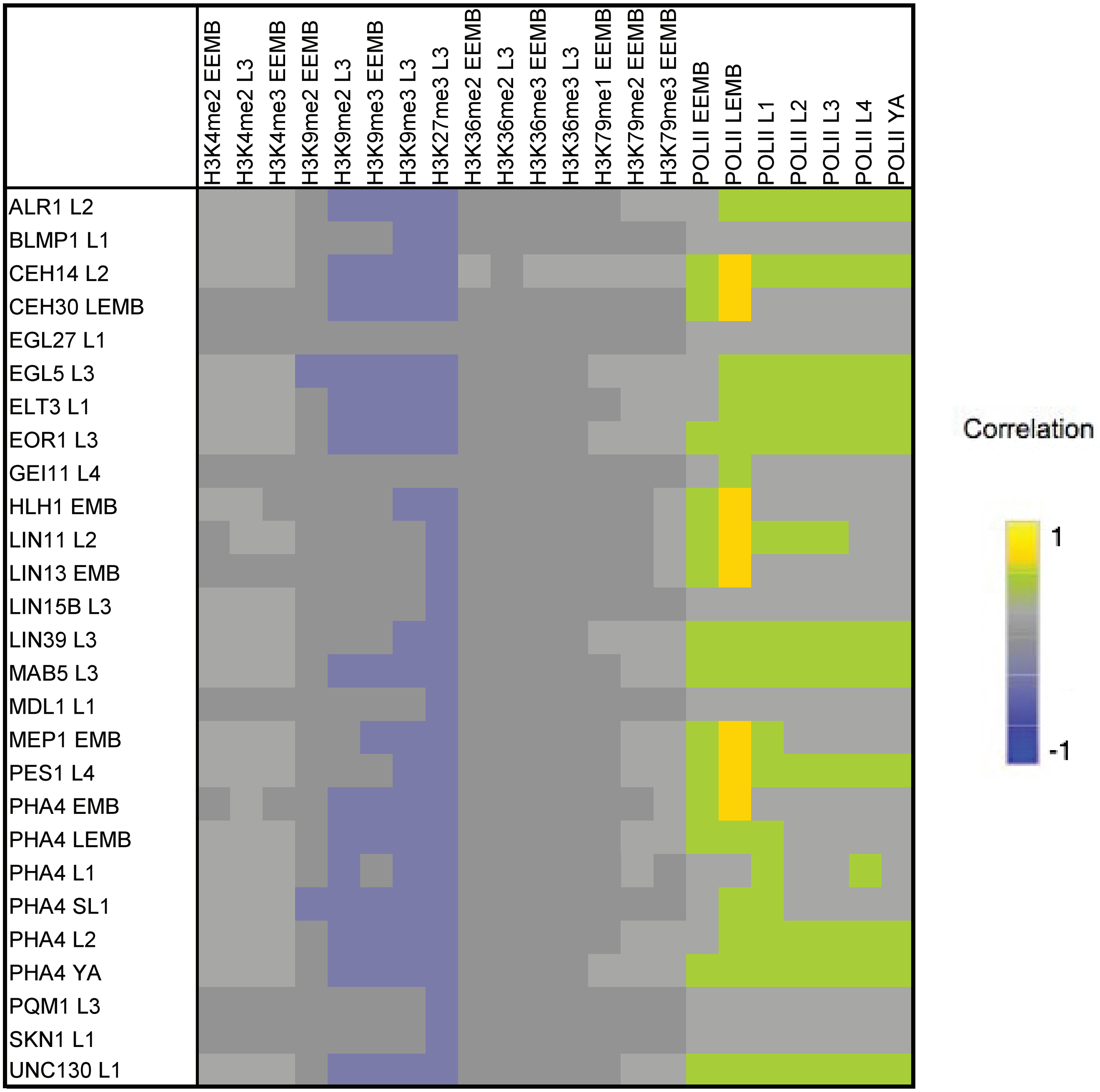

H3K4me2 EEMB
H3K4me2 L3
H3K4me3 EEMB
H3K9me2 EEMB
H3K9me2 L3
H3K9me3 EEMB
H3K9me3 L3
H3K27me3 L3
H3K36me2 EEMB
H3K36me2 L3
H3K36me3 EEMB
H3K36me3 L3
H3K79me1 EEMB
H3K79me2 EEMB
H3K79me3 EEMB
POLII EEMB
POLII LEMB
POLII L1
POLII L2
POLII L3
POLII L4
POLII YA
ALR1 L2
BLMP1 L1
CEH14 L2
CEH30 LEMB
EGL27 L1
EGL5 L3
ELT3 L1
EOR1 L3
GEI11 L4
HLH1 EMB
LIN11 L2
LIN13 EMB
LIN15B L3
LIN39 L3
MAB5 L3
MDL1 L1
MEP1 EMB
PES1 L4
PHA4 EMB
PHA4 LEMB
PHA4 L1
PHA4 SL1
PHA4 L2
PHA4 YA
PQM1 L3
SKN1 L1
UNC130 L1
Correlation
1
-1

Fig. S38A

Predictors Binding experiments	H3K4me2 EE	H3K4me2 L3	H3K4me3 EE	H3K9me2 EE	H3K9me2 L3	H3K9me3 EE	H3K9me3 L3	H3K27me3 L3	H3K36me2 EE	H3K36me2 L3	H3K36me3 EE	H3K36me3 L3	H3K79me1 EE	H3K79me2 EE	H3K79me3 EE	RNA POLII EE	RNA POLII LE	RNA POLII L1	RNA POLII L2	RNA POLII L3	RNA POLII L4	RNA POLII YA	Histone marks	All RNA POLII	All features
ALR1 L2	0.66	0.76	0.67	0.61	0.64	0.65	0.70	0.67	0.57	0.57	0.49	0.62	0.56	0.61	0.58	0.65	0.77	0.82	0.83	0.82	0.80	0.80	0.87	0.85	0.90
BLMP1 L1	0.66	0.77	0.65	0.62	0.57	0.63	0.70	0.63	0.57	0.59	0.53	0.61	0.51	0.61	0.57	0.64	0.76	0.81	0.86	0.84	0.83	0.83	0.86	0.88	0.91
CEH14 L2	0.86	0.89	0.84	0.58	0.67	0.63	0.75	0.88	0.70	0.53	0.65	0.60	0.76	0.82	0.80	0.84	0.90	0.94	0.94	0.93	0.93	0.91	0.95	0.95	0.96
CEH30 LE	0.85	0.83	0.82	0.58	0.64	0.54	0.75	0.86	0.59	0.55	0.56	0.60	0.67	0.77	0.74	0.83	0.94	0.95	0.95	0.94	0.94	0.92	0.96	0.97	0.98
EGL27 L1	0.72	0.78	0.70	0.50	0.64	0.59	0.76	0.56	0.55	0.61	0.51	0.68	0.57	0.65	0.63	0.74	0.79	0.86	0.90	0.89	0.84	0.86	0.89	0.91	0.94
EGL5 L3	0.79	0.81	0.72	0.59	0.64	0.57	0.80	0.78	0.56	0.70	0.59	0.75	0.58	0.60	0.71	0.79	0.92	0.96	0.96	0.96	0.96	0.94	0.95	0.98	0.98
ELT3 L1	0.68	0.72	0.65	0.50	0.64	0.61	0.67	0.69	0.57	0.58	0.56	0.64	0.57	0.49	0.61	0.66	0.65	0.90	0.90	0.89	0.89	0.85	0.86	0.93	0.93
EOR1 L3	0.79	0.85	0.76	0.57	0.64	0.52	0.69	0.84	0.57	0.53	0.58	0.61	0.66	0.63	0.70	0.78	0.83	0.96	0.95	0.95	0.94	0.94	0.94	0.97	0.98
GEI11 L4	0.69	0.66	0.68	0.61	0.64	0.57	0.58	0.60	0.50	0.59	0.56	0.66	0.48	0.54	0.57	0.71	0.76	0.81	0.83	0.84	0.82	0.80	0.91	0.85	0.93
HLH1 MxE	0.82	0.75	0.70	0.50	0.66	0.66	0.76	0.72	0.59	0.63	0.60	0.67	0.59	0.68	0.67	0.79	0.89	0.83	0.86	0.85	0.84	0.83	0.91	0.91	0.95
LIN11 L2	0.74	0.80	0.83	0.57	0.64	0.56	0.73	0.82	0.67	0.62	0.64	0.61	0.62	0.74	0.73	0.79	0.90	0.96	0.96	0.95	0.94	0.93	0.93	0.97	0.96
LIN13 MxE	0.89	0.86	0.88	0.59	0.55	0.56	0.72	0.84	0.58	0.59	0.56	0.61	0.69	0.80	0.74	0.81	0.92	0.93	0.94	0.93	0.93	0.85	0.96	0.97	0.98
LIN15B L3	0.81	0.87	0.84	0.58	0.60	0.57	0.67	0.85	0.69	0.62	0.62	0.56	0.63	0.77	0.70	0.77	0.88	0.93	0.94	0.93	0.90	0.90	0.94	0.96	0.98
LIN39 L3	0.72	0.83	0.64	0.52	0.64	0.64	0.75	0.75	0.53	0.59	0.57	0.64	0.57	0.55	0.63	0.75	0.86	0.85	0.87	0.88	0.86	0.85	0.90	0.92	0.94
MAB5 L3	0.77	0.80	0.75	0.54	0.59	0.48	0.74	0.81	0.54	0.58	0.54	0.65	0.63	0.69	0.67	0.76	0.84	0.92	0.92	0.91	0.91	0.90	0.90	0.94	0.95
MDL1 L1	0.78	0.69	0.61	0.51	0.62	0.63	0.66	0.78	0.51	0.55	0.57	0.62	0.63	0.68	0.65	0.73	0.85	0.95	0.93	0.92	0.93	0.91	0.92	0.96	0.97
MEP1 MxE	0.86	0.86	0.83	0.58	0.63	0.60	0.72	0.86	0.64	0.54	0.53	0.60	0.63	0.76	0.74	0.80	0.91	0.91	0.93	0.92	0.93	0.91	0.96	0.96	0.98
PES1 L4	0.77	0.84	0.78	0.49	0.62	0.57	0.65	0.76	0.66	0.60	0.62	0.55	0.58	0.72	0.69	0.74	0.84	0.91	0.89	0.91	0.91	0.89	0.95	0.93	0.97
PHA4 MxE	0.83	0.85	0.75	0.52	0.64	0.64	0.70	0.76	0.50	0.55	0.50	0.60	0.65	0.72	0.68	0.77	0.89	0.88	0.89	0.89	0.88	0.86	0.93	0.93	0.94
PHA4 LE	0.65	0.79	0.66	0.60	0.54	0.62	0.70	0.73	0.51	0.58	0.54	0.50	0.58	0.63	0.55	0.68	0.78	0.85	0.85	0.83	0.82	0.80	0.87	0.86	0.90
PHA4 L1	0.73	0.80	0.69	0.59	0.60	0.59	0.71	0.77	0.58	0.56	0.55	0.63	0.59	0.66	0.62	0.66	0.83	0.87	0.90	0.86	0.89	0.86	0.90	0.92	0.92
PHA4 stvL1	0.80	0.76	0.75	0.57	0.54	0.62	0.70	0.79	0.60	0.54	0.57	0.61	0.64	0.70	0.60	0.74	0.87	0.89	0.85	0.88	0.87	0.84	0.91	0.92	0.95
PHA4 L2	0.61	0.78	0.60	0.48	0.59	0.62	0.66	0.69	0.55	0.59	0.52	0.65	0.57	0.63	0.60	0.68	0.74	0.85	0.84	0.83	0.83	0.82	0.86	0.87	0.91
PHA4 YA	0.73	0.83	0.72	0.52	0.68	0.61	0.72	0.79	0.57	0.61	0.57	0.63	0.58	0.69	0.68	0.78	0.86	0.96	0.96	0.95	0.95	0.95	0.91	0.97	0.97
PQM1 L3	0.58	0.67	0.59	0.57	0.56	0.53	0.63	0.63	0.60	0.61	0.59	0.64	0.49	0.57	0.59	0.60	0.67	0.82	0.81	0.80	0.81	0.76	0.83	0.86	0.89
SKN1 L1	0.80	0.85	0.72	0.58	0.58	0.60	0.71	0.83	0.59	0.53	0.59	0.60	0.56	0.70	0.60	0.73	0.85	0.93	0.94	0.92	0.91	0.90	0.94	0.95	0.96
UNC130 L1	0.75	0.73	0.72	0.57	0.64	0.63	0.73	0.71	0.57	0.59	0.56	0.64	0.61	0.63	0.63	0.71	0.82	0.95	0.94	0.93	0.91	0.92	0.90	0.96	0.96
HOT (core)	0.84	0.79	0.88	0.61	0.70	0.64	0.77	0.89	0.51	0.59	0.57	0.69	0.66	0.81	0.80	0.87	0.97	0.99	0.99	0.99	0.98	0.97	0.98	0.99	1.00
HOT (exteneded)	0.86	0.88	0.85	0.54	0.66	0.58	0.75	0.88	0.68	0.61	0.64	0.64	0.75	0.84	0.84	0.85	0.89	0.94	0.93	0.92	0.92	0.93	0.96	0.95	0.97

Fig. S38B

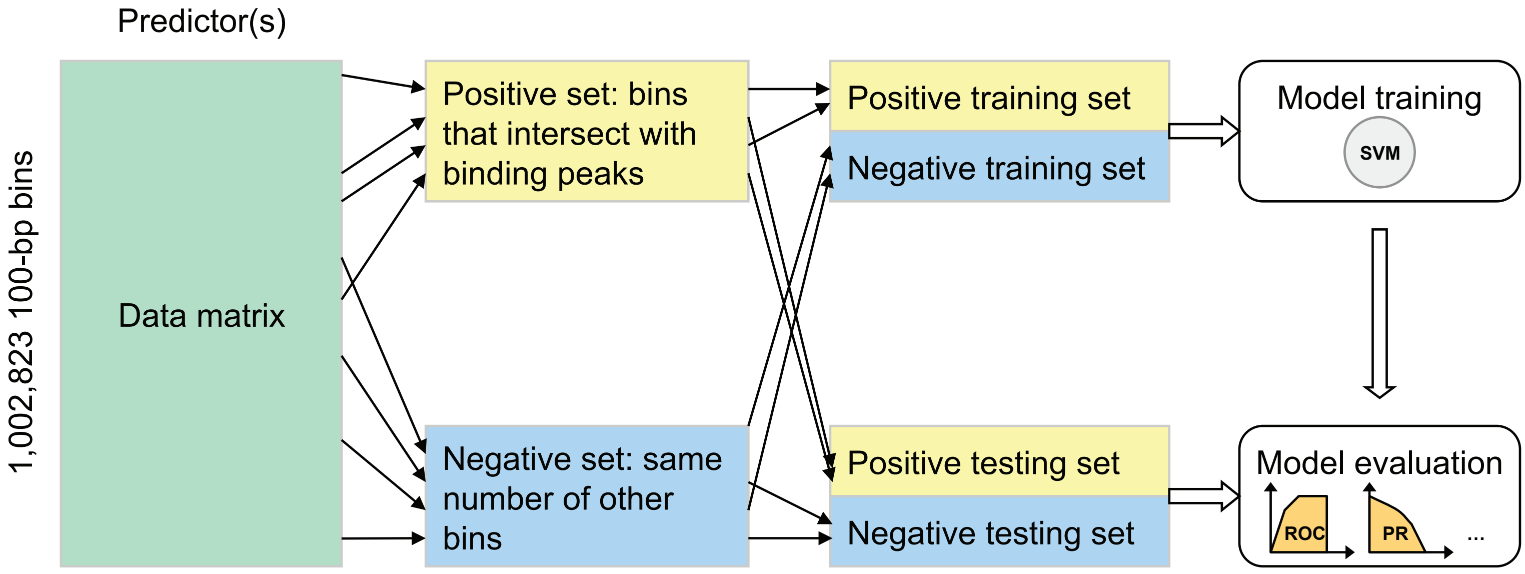

Predictor(s)
1,002,823 100-bp bins
Data matrix
Positive set: bins that intersect with binding peaks
Negative set: same number of other bins
Positive training set
Negative training set
Positive testing set
Negative testing set
Model training
SVM
Model evaluation
ROC
PR
...

Fig. S39

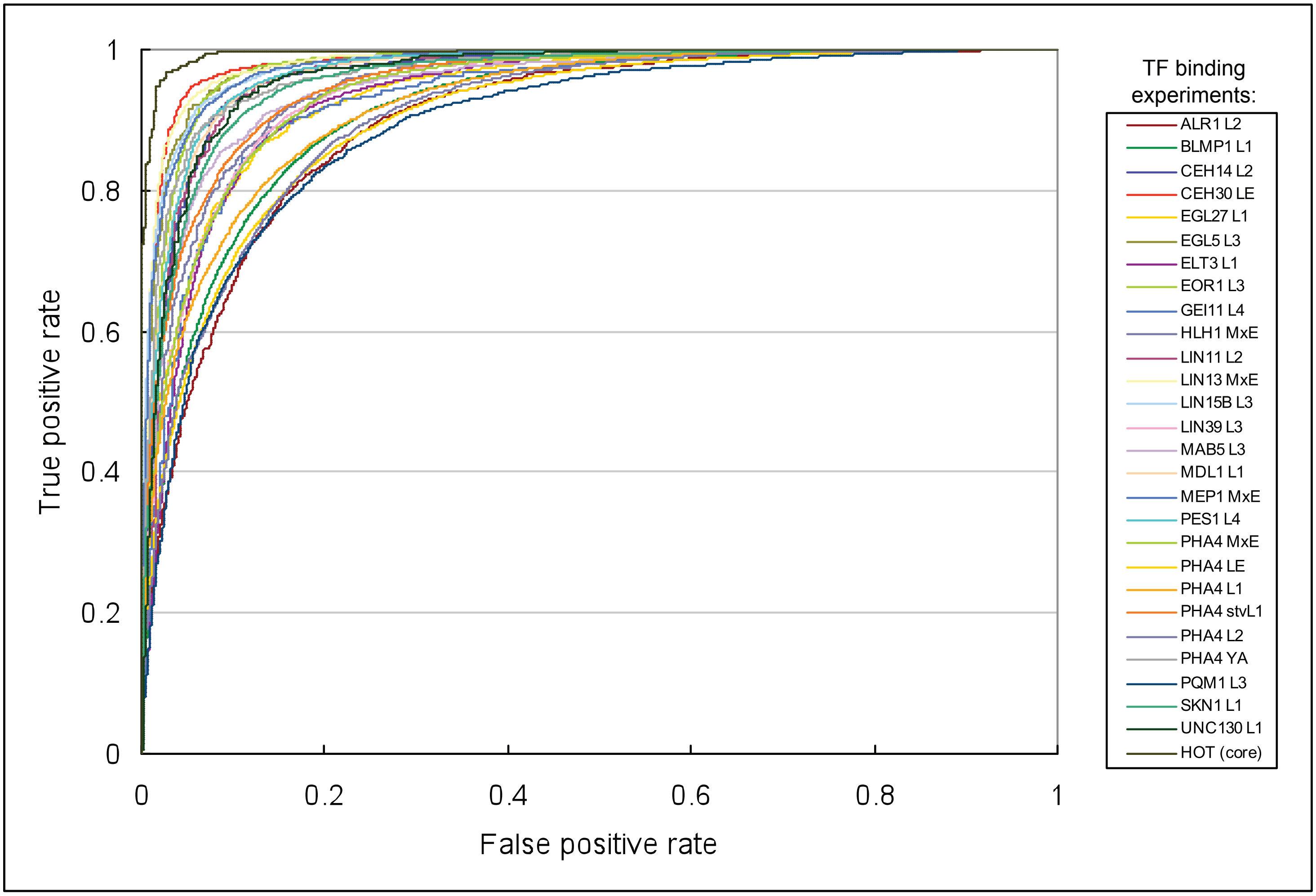

1
0.8
0.6
0.4
0.2
0
True positive rate
0
0.2
0.4
0.6
0.8
1
False positive rate
TF binding experiments:
ALR1 L2
BLMP1 L1
CEH14 L2
CEH30 LE
EGL27 L1
EGL5 L3
ELT3 L1
EOR1 L3
GEI11 L4
HLH1 MxE
LIN11 L2
LIN13 MxE
LIN15B L3
LIN39 L3
MAB5 L3
MDL1 L1
MEP1 MxE
PES1 L4
PHA4 MxE
PHA4 LE
PHA4 L1
PHA4 stvL1
PHA4 L2
PHA4 YA
PQM1 L3
SKN1 L1
UNC130 L1
HOT (core)

Fig. S40A

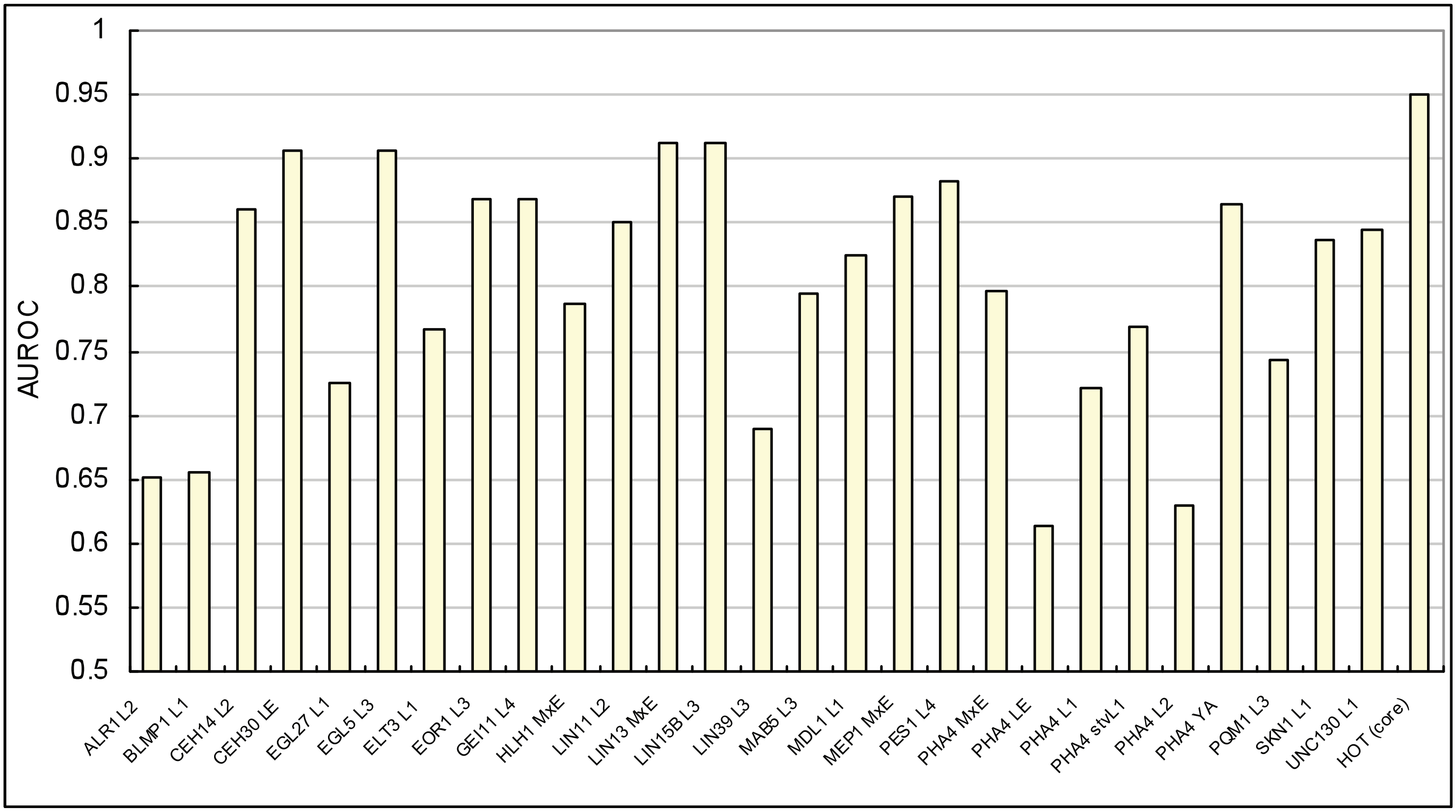

1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
AUROC
ALR1 L2
BLMP1 L1
CEH14 L2
CEH30 LE
EGL27 L1
EGL5 L3
ELT3 L1
EOR1 L3
GEI11 L4
HLH1 MxE
LIN11 L2
LIN13 MxE
LIN15B L3
LIN39 L3
MAB5 L3
MDL1 L1
MEP1 MxE
PES1 L4
PHA4 MxE
PHA4 LE
PHA4 L1
PHA4 stvL1
PHA4 L2
PHA4 YA
PQM1 L3
SKN1 L1
UNC130 L1
HOT (core)

Fig. S40B

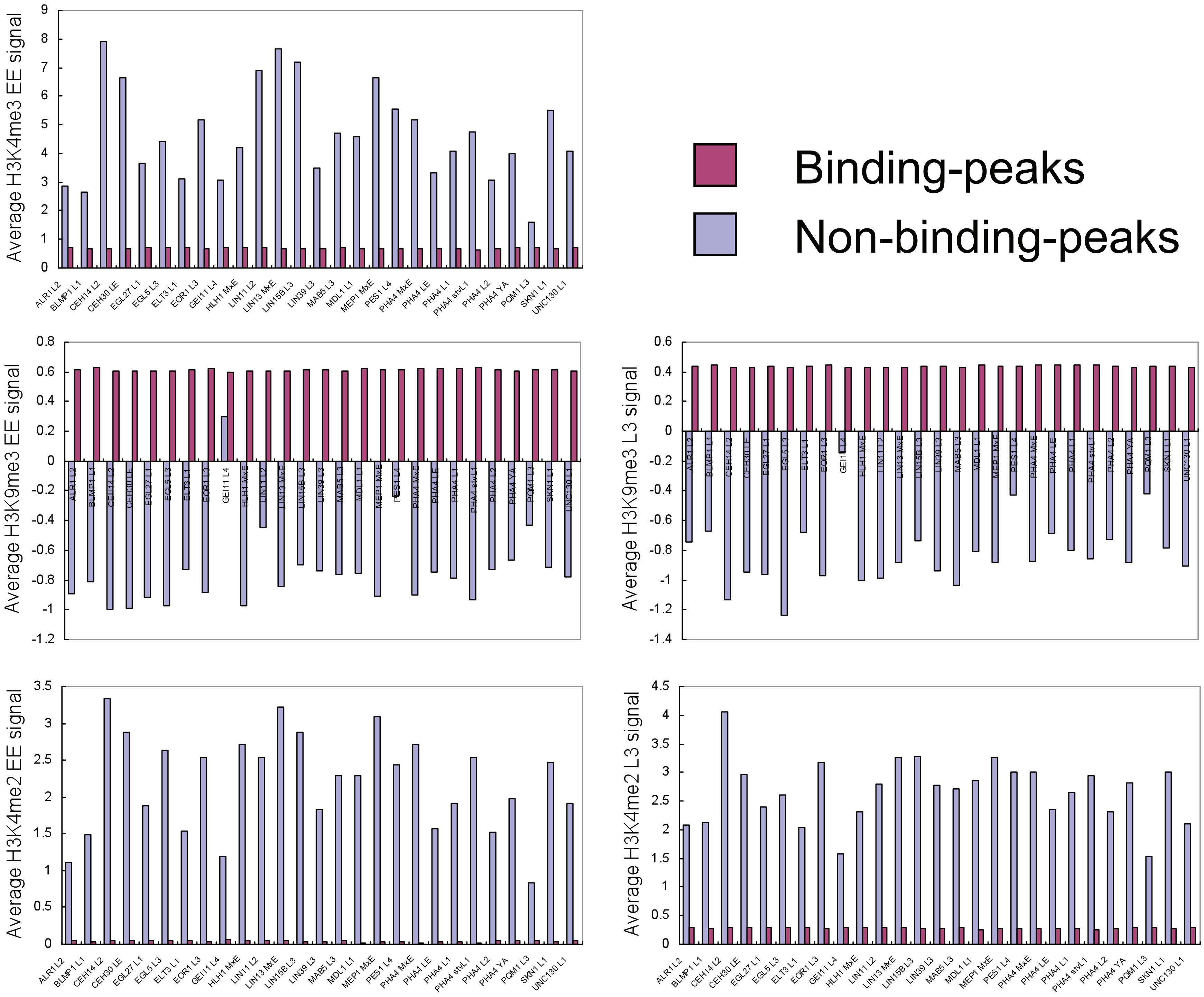

Average H3K4me3 EE signal
9
8
7
6
5
4
3
2
1
0
Binding-peaks
Non-binding-peaks
Average H3K9me3 EE signal
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
Average H3K9me3 L3 signal
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
Average H3K4me2 EE signal
3.5
3
2.5
2
1.5
1
0.5
0
Average H3K4me2 L3 signal
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
ALR1 L2
BLMP1 L1
CEH14 L2
CEH30 LE
EGL27 L1
EGL5 L3
ELT3 L1
EOR1 L3
GEI11 L4
HLH1 MxE
LIN11 L2
LIN13 MxE
LIN15B L3
LIN39 L3
MAB5 L3
MDL1 L1
MEP1 MxE
PES1 L4
PHA4 MxE
PHA4 LE
PHA4 L1
PHA4 stvL1
PHA4 L2
PHA4 YA
PQM1 L3
SKN1 L1
UNC130 L1

Fig. S41

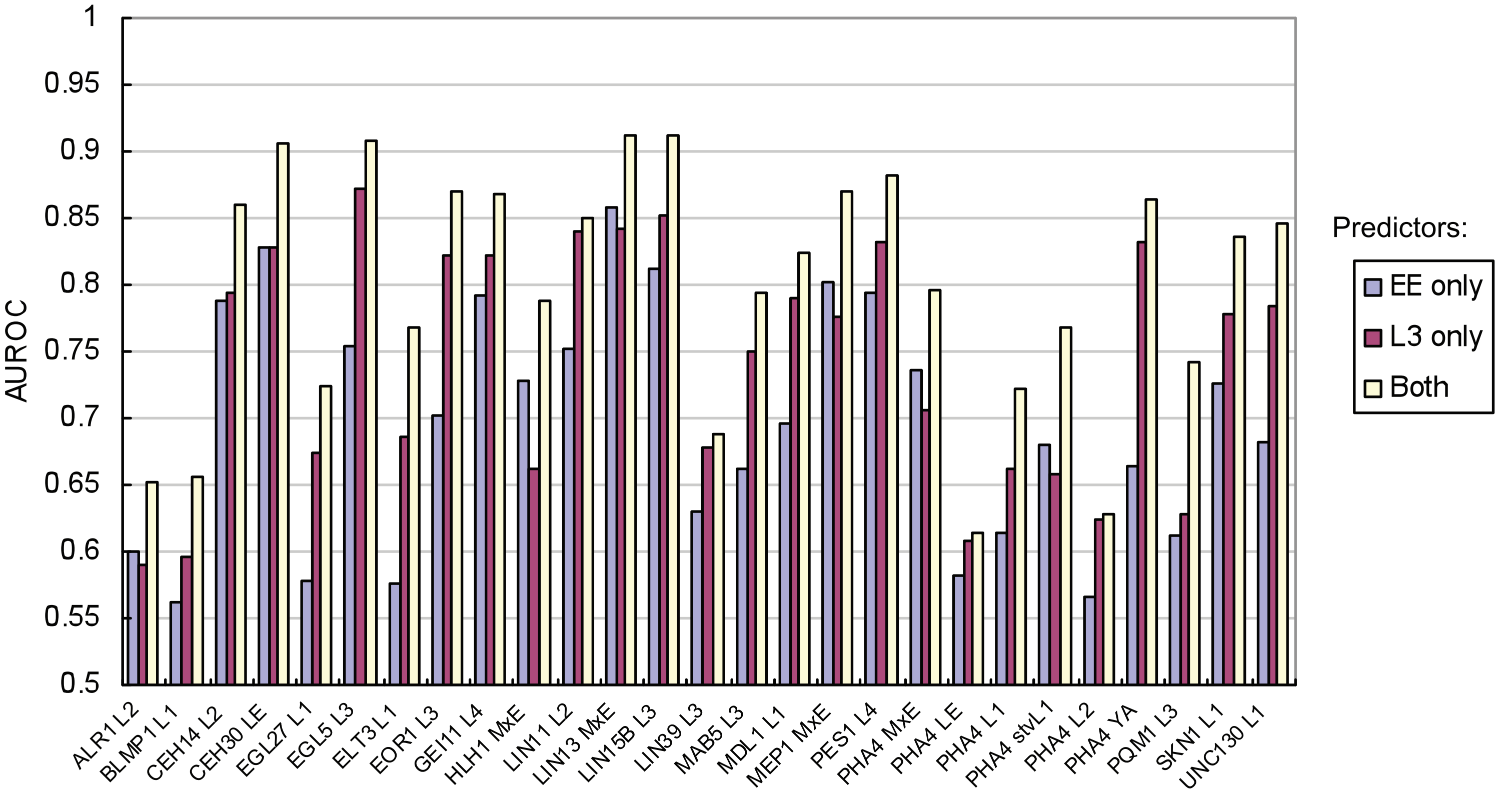
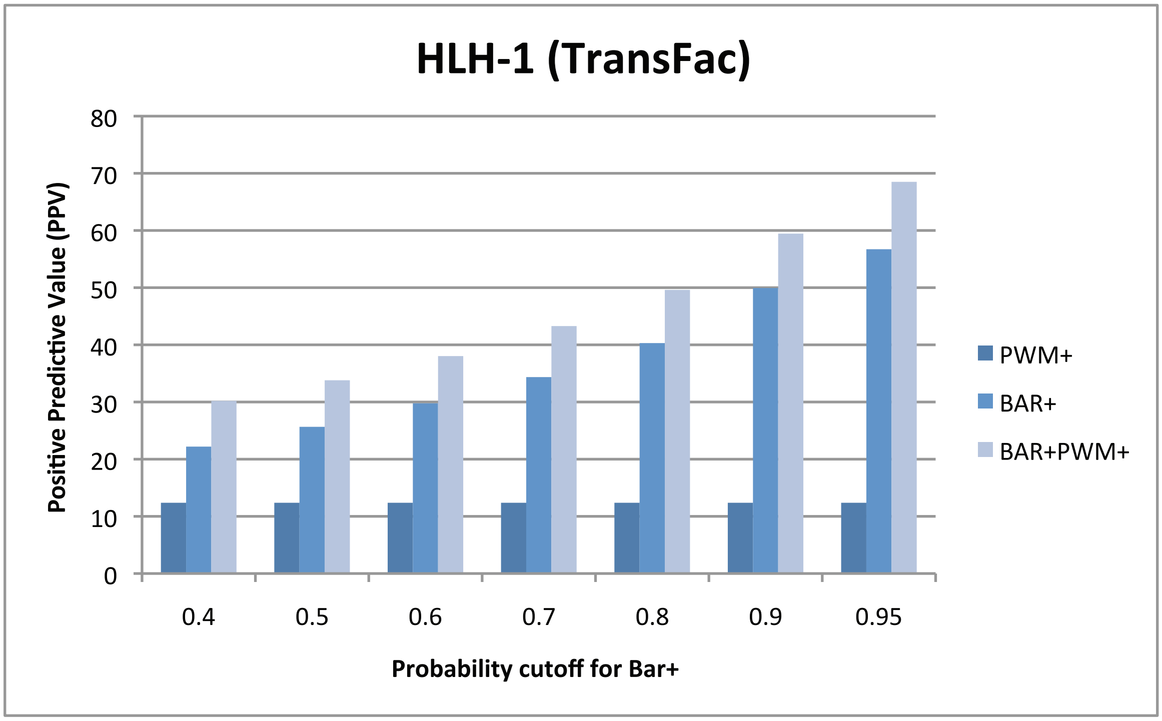

AUROC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
ALR1 L2
BLMP1 L1
CEH14 L2
CEH30 LE
EGL27 L1
EGL5 L3
ELT3 L1
EOR1 L3
GEI11 L4
HLH1 MxE
LIN11 L2
LIN13 MxE
LIN15B L3
LIN39 L3
MAB5 L3
MDL1 L1
MEP1 MxE
PES1 L4
PHA4 MxE
PHA4 LE
PHA4 L1
PHA4 stvL1
PHA4 L2
PHA4 YA
PQM1 L3
SKN1 L1
UNC130 L1
Predictors:
EE only
L3 only
Both

Fig. S42


HLH-1 (TransFac)
Positive Predictive Value (PPV)
80
70
60
50
40
30
20
10
0
0.4
0.5
0.6
0.7
0.8
0.9
0.95
Probability cutoff for Bar+
PWM+
BAR+
BAR+PWM+

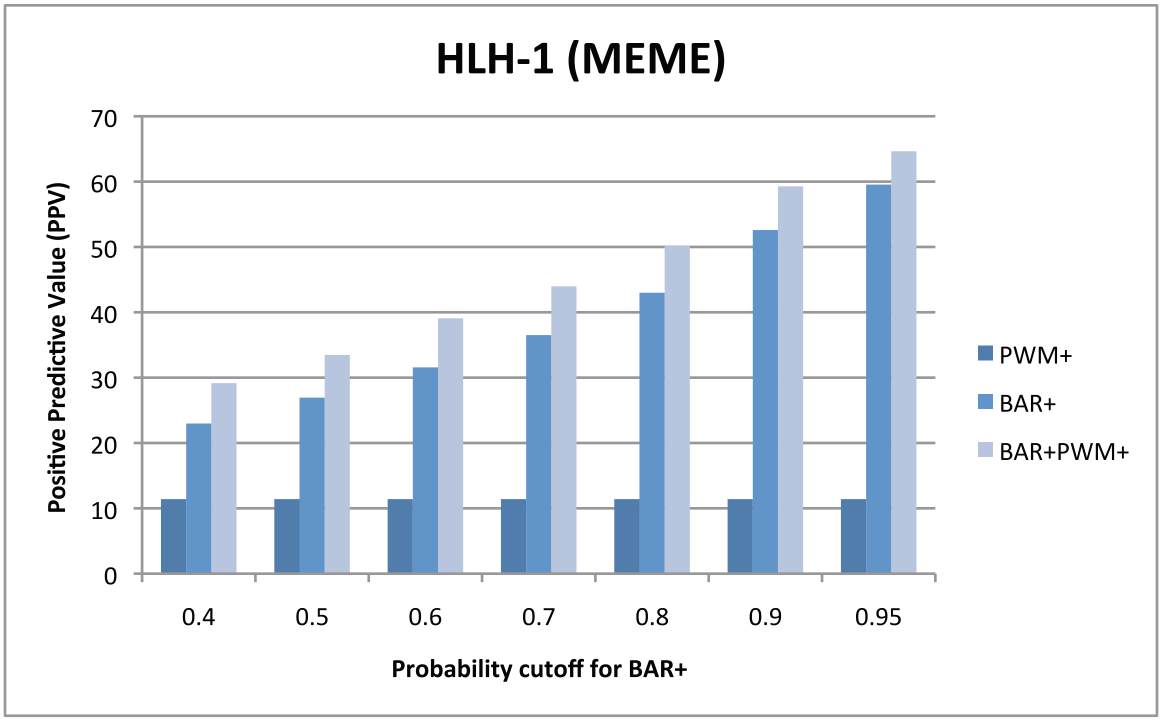

HLH-1 (MEME)
Positive Predictive Value (PPV)
70
60
50
40
30
20
10
0
0.4
0.5
0.6
0.7
0.8
0.9
0.95
Probability cutoff for BAR+
PWM+
BAR+
BAR+PWM+

Fig. S43

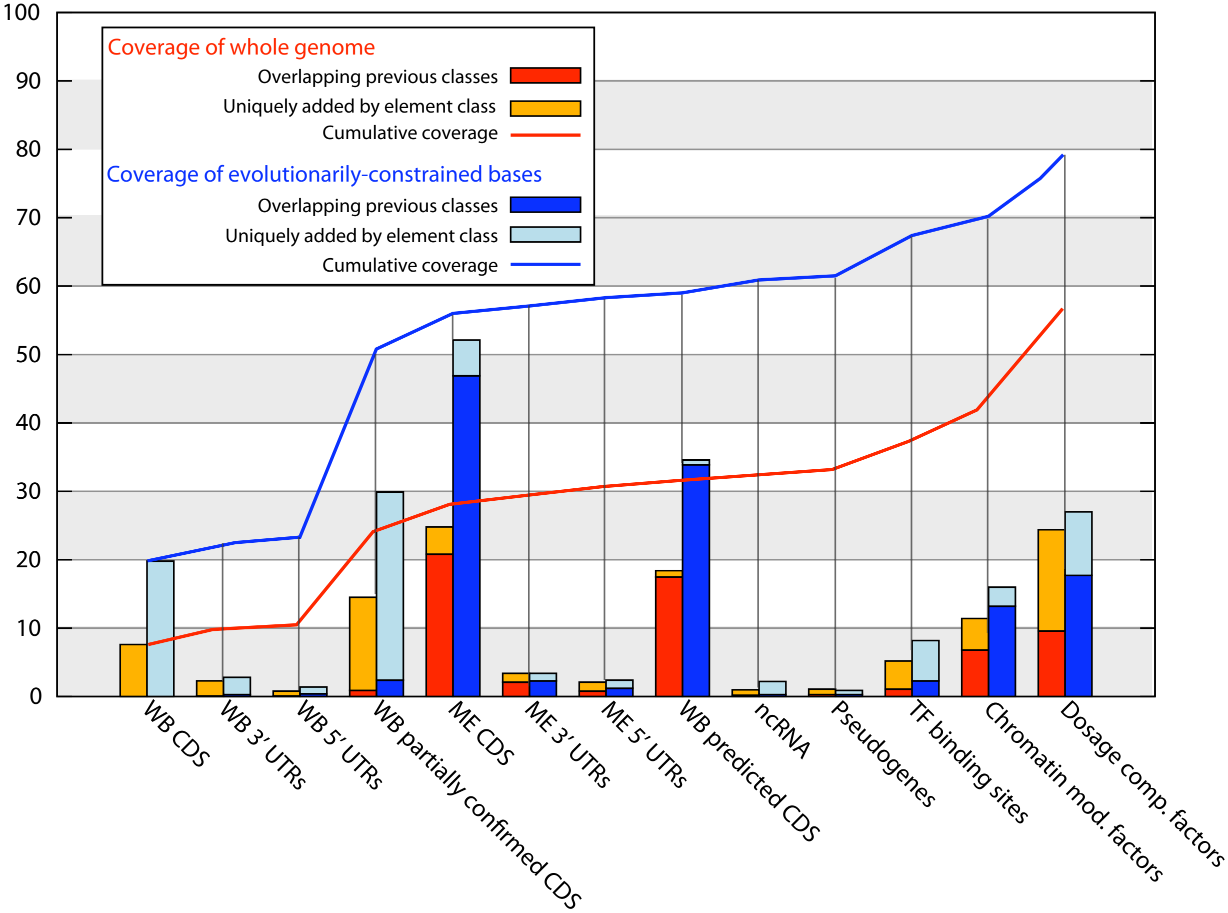

100
90
80
70
60
50
40
30
20
10
0
Coverage of whole genome
Overlapping previous classes
Uniquely added by element class
Cumulative coverage
Coverage of evolutionarily-constrained bases
Overlapping previous classes
Uniquely added by element class
Cumulative coverage
WB CDS
WB 3' UTRs
WB 5' UTRs
WB partially confirmed CDS
ME CDS
ME 3' UTRs
ME 5' UTRs
WB predicted CDS
ncRNA
Pseudogenes
TF binding sites
Chromatin mod. factors
Dosage comp. factors

Fig. S44

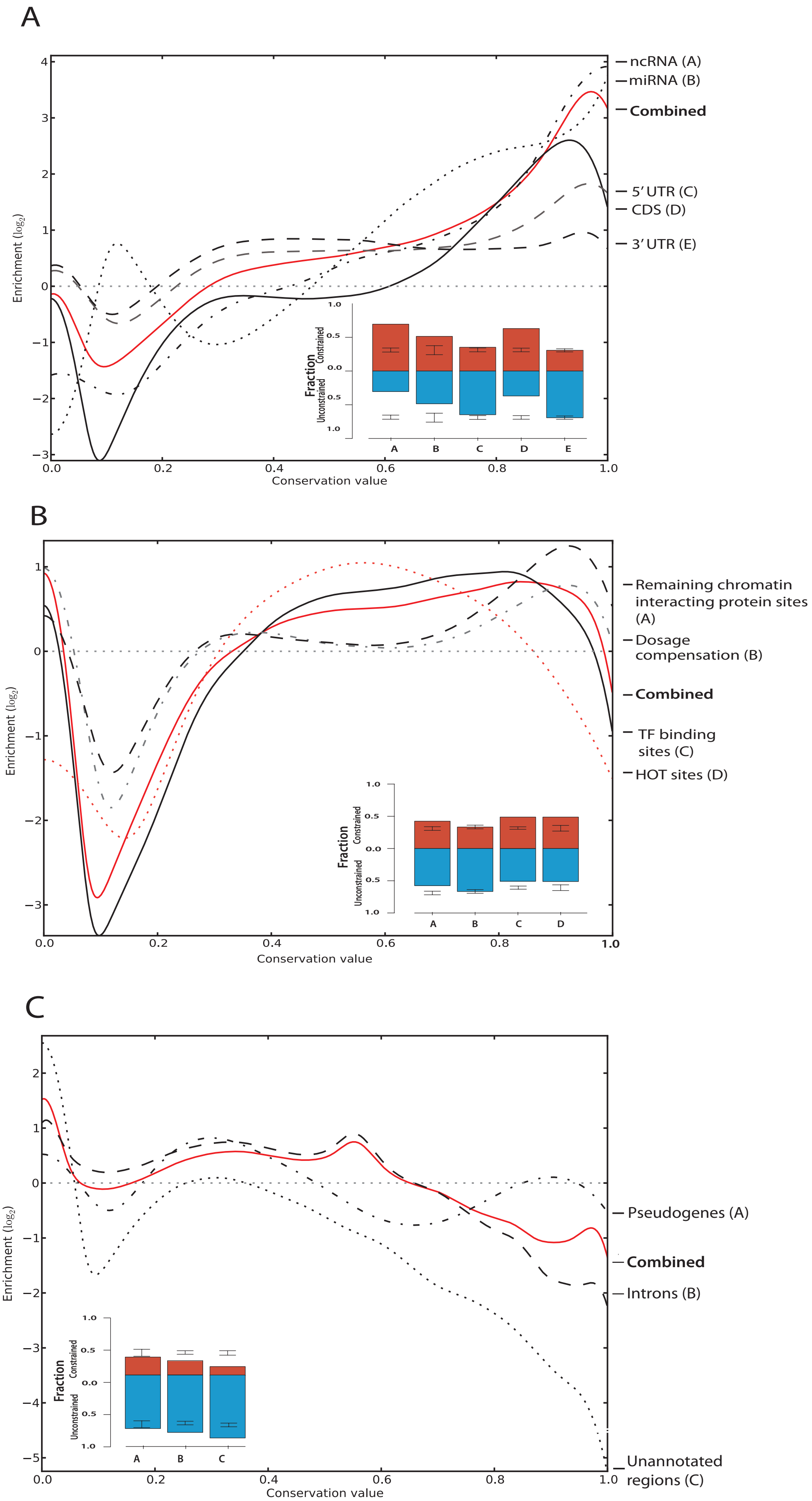

A
ncRNA (A)
miRNA (B)
Combined
5′ UTR (C)
CDS (D)
3′ UTR (E)
Enrichment (log2)
4
3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0
Conservation value
Fraction
Constrained
Unconstrained
1.0
0.5
0.0
0.5
1.0
A
B
C
D
E
B
Remaining chromatin interacting protein sites (A)
Dosage compensation (B)
Combined
TF binding sites (C)
HOT sites (D)
Enrichment (log2)
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0
Conservation value
Fraction
Constrained
Unconstrained
1.0
0.5
0.0
0.5
1.0
A
B
C
D
C
Pseudogenes (A)
Combined
Introns (B)
Unannotated regions (C)
Enrichment (log2)
2
1
0
−1
−2
−3
−4
−5
0.0
0.2
0.4
0.6
0.8
1.0
Conservation value
Fraction
Constrained
Unconstrained
1.0
0.5
0.0
0.5
1.0
A
B
C

Fig. S45

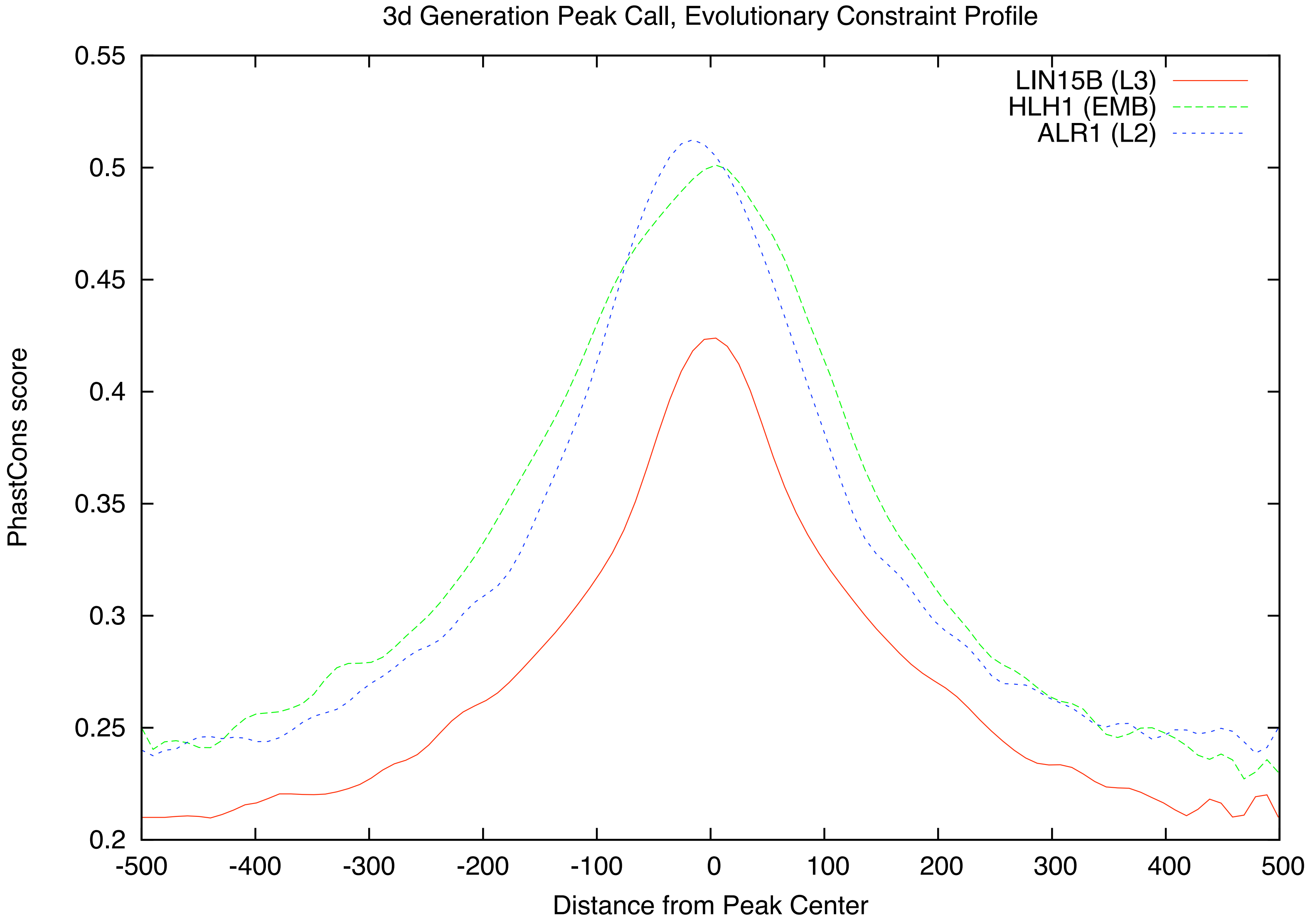

3d Generation Peak Call, Evolutionary Constraint Profile
LIN15B (L3)
HLH1 (EMB)
ALR1 (L2)
PhastCons score
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
-500
-400
-300
-200
-100
0
100
200
300
400
500
Distance from Peak Center

Fig. S46

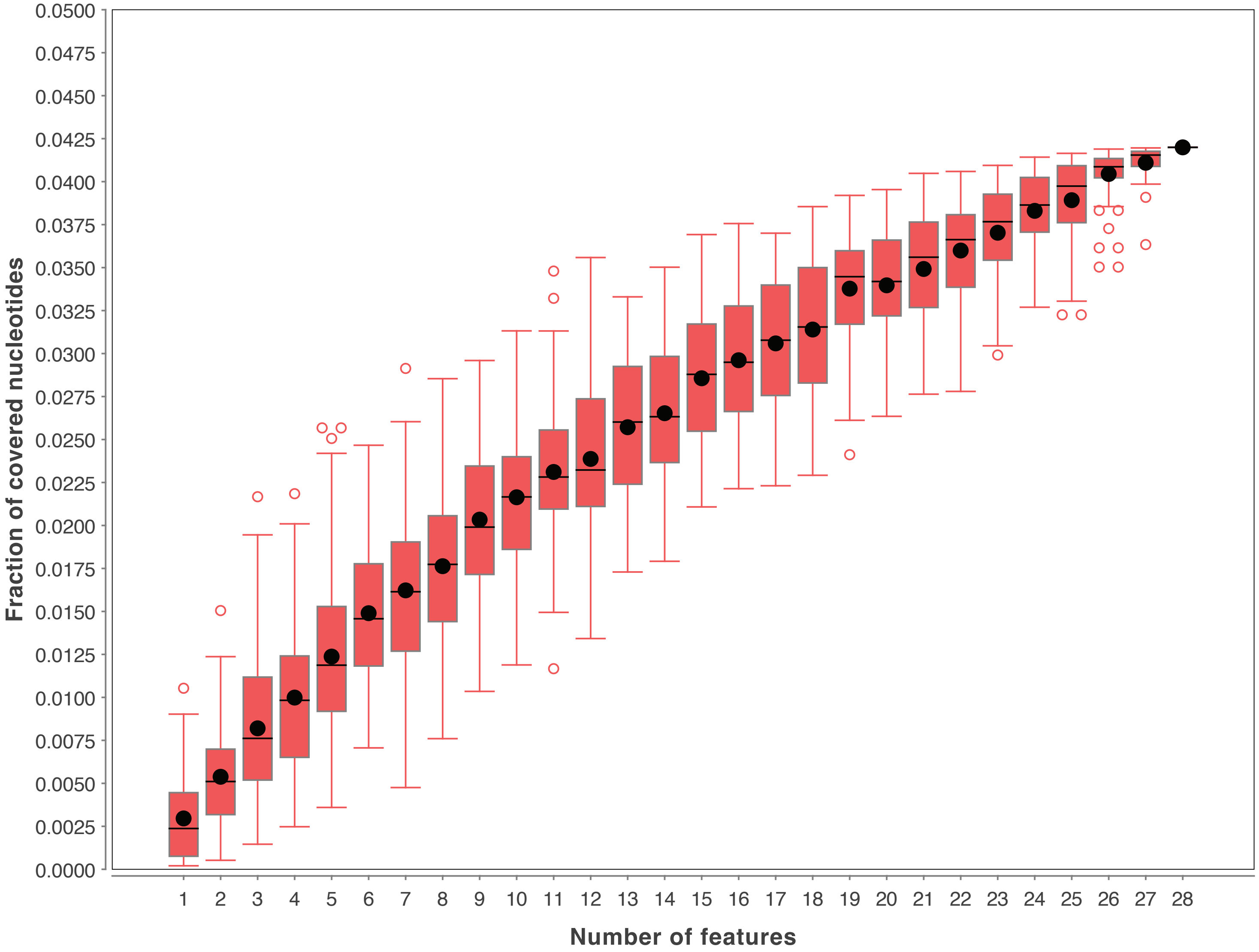
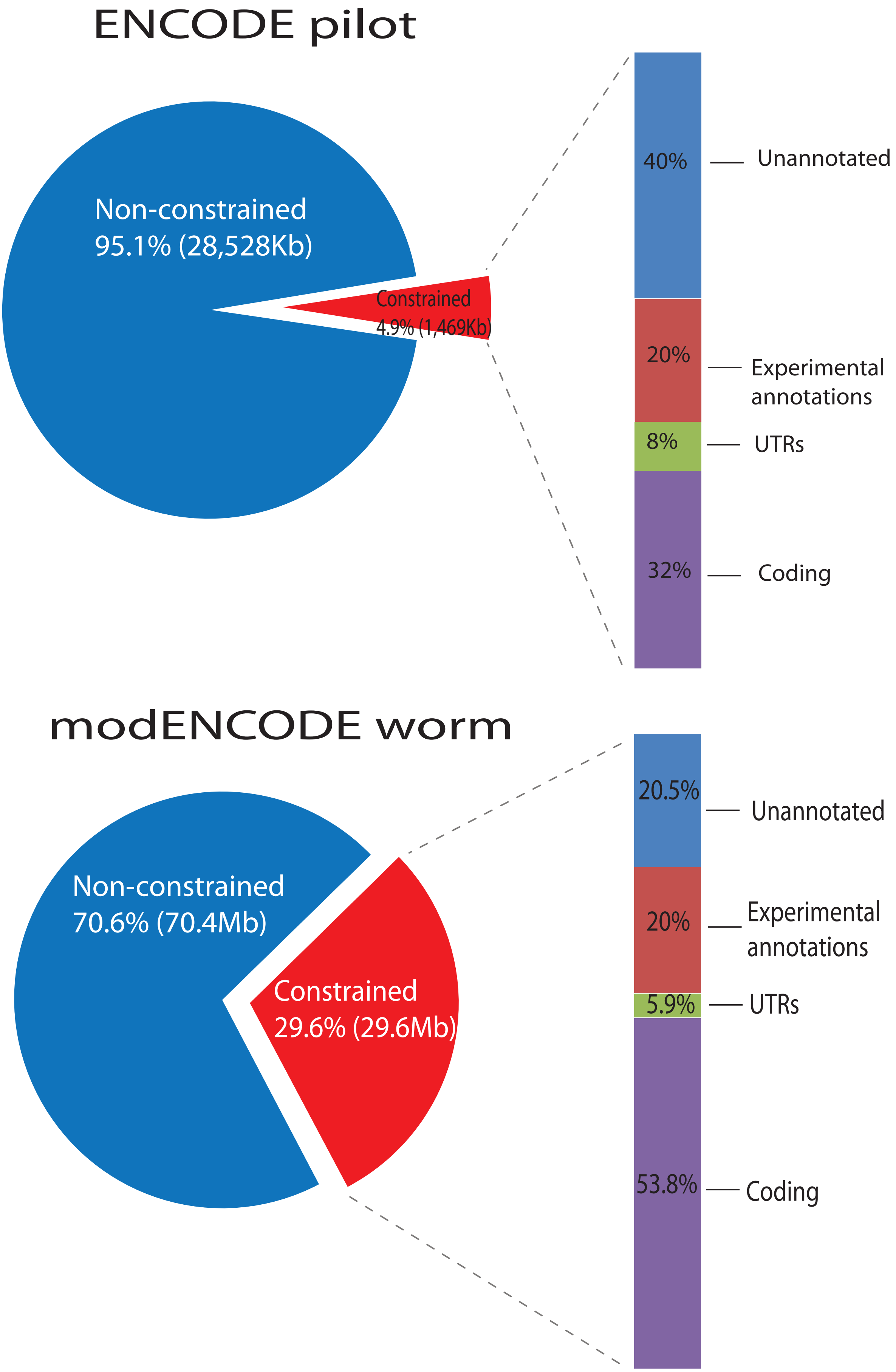

0.0500
0.0475
0.0450
0.0425
0.0400
0.0375
0.0350
0.0325
0.0300
0.0275
0.0250
0.0225
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Fraction of covered nucleotides
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Number of features

Fig. S47



ENCODE pilot

modENCODE worm

Fig. S48

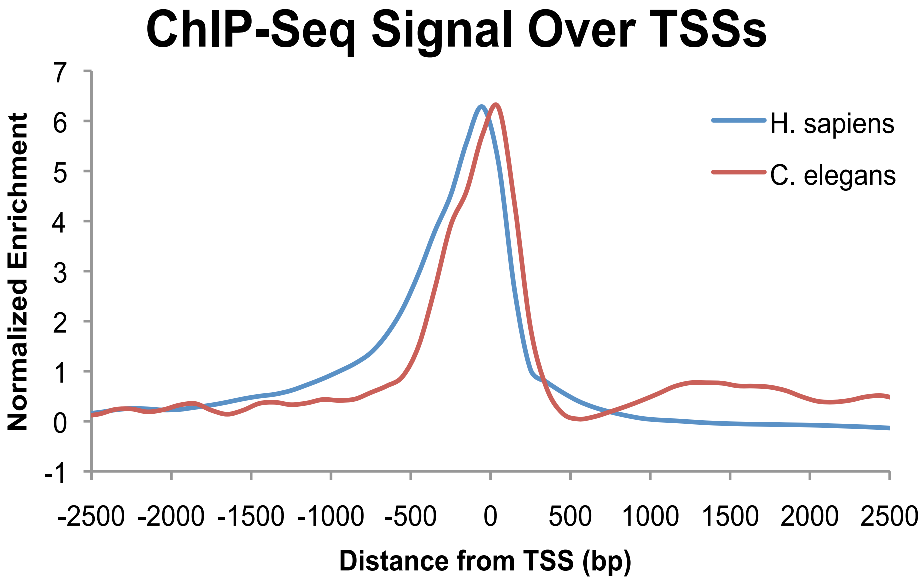

ChIP-Seq Signal Over TSSs
Normalized Enrichment
7
6
5
4
3
2
1
0
-1
-2500
-2000
-1500
-1000
-500
0
500
1000
1500
2000
2500
Distance from TSS (bp)
H. sapiens
C. elegans

Fig. S49

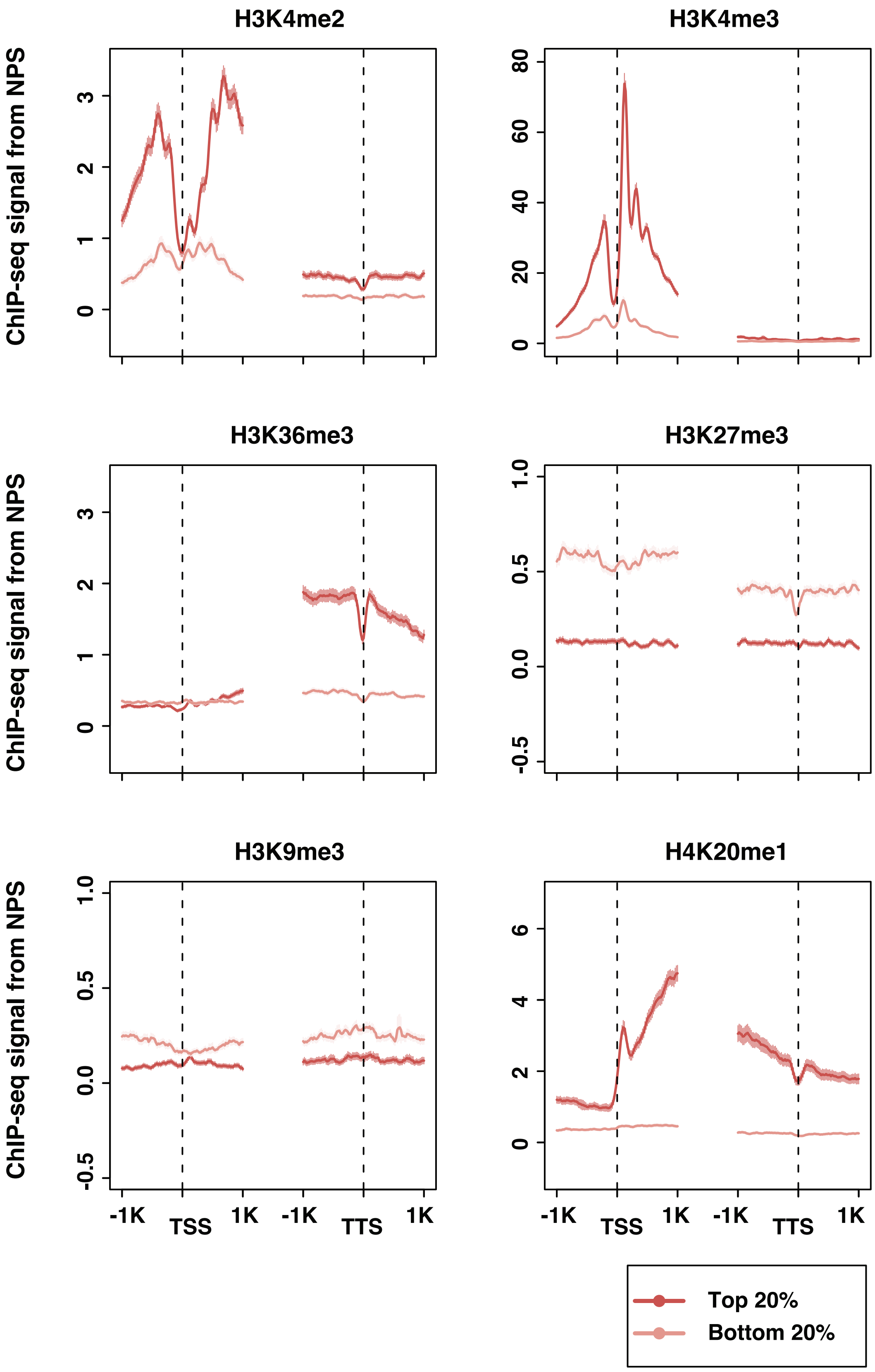

H3K4me2
H3K4me3
H3K36me3
H3K27me3
H3K9me3
H4K20me1
ChIP-seq signal from NPS
0
1
2
3
20
40
60
80
-0.5
0.0
0.5
1.0
4
6
-1K
TSS
1K
-1K
TTS
1K
Top 20%
Bottom 20%

Fig. S50